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.

ISSN 2395-8618

***Index Terms*— Natural language processing, topic modeling, clustering, study programs, NMF, automatic curriculum analysis.**

I. Introducción

Las instituciones de educación superior generan y mantienen grandes volúmenes de documentación curricular: programas de estudio, guías de asignatura, perfiles de egreso y documentos normativos. Estos repositorios contienen información valiosa sobre qué se enseña, cómo se estructura la formación profesional y qué relaciones existen entre asignaturas. Sin embargo, el análisis de estos documentos se realiza predominantemente de forma manual, mediante la lectura individual por parte de comités académicos que, debido a limitaciones de tiempo y alcance, suelen enfocarse en subconjuntos reducidos del plan de estudios [12].

Esta situación genera dificultades concretas: identificar qué áreas temáticas predominan en un plan de estudios, detectar solapamientos entre asignaturas, localizar vacíos conceptuales o evaluar la coherencia global de un programa formativo son tareas que requieren un análisis comparativo exhaustivo difícil de sostener manualmente cuando el volumen de documentos es considerable.

Las revisiones sistemáticas sobre inteligencia artificial (IA) en educación superior revelan que la investigación se ha concentrado mayoritariamente en sistemas tutoriales inteligentes, predicción de deserción y evaluación automatizada. Zawacki-Richter et al. [15] encontraron que solo el 8.2 % de los 146 estudios revisados abordaban el análisis de documentos institucionales o curriculares. Wang et al. [13] actualizaron este panorama con 361 publicaciones entre 2017 y 2023, confirmando que menos del 5 % se relaciona con análisis de programas de estudio, mientras que temas como la IA generativa concentran más del 40 % de las publicaciones recientes. Esta distribución evidencia un vacío significativo en herramientas de PLN orientadas al análisis curricular. García-Peñalvo [5] observa que, pese al creciente interés por la IA en educación, las aplicaciones se concentran en interacciones directas con el estudiante, no en la gestión documental institucional.

En el contexto mexicano, el problema adquiere una dimensión particular. El sistema de institutos tecnológicos opera 254 campus distribuidos en todo el territorio nacional, atendiendo a más de 600,000 estudiantes en 41 programas de ingeniería [12]. Los planes de estudio son de alcance nacional, con ciclos de actualización de 3 a 5 años que involucran academias distribuidas en todos los campus. Cada plan contiene entre 40 y 55 asignaturas, lo que implica la gestión de entre 10,000 y 13,000 programas de estudio a nivel sistema. Las revisiones curriculares dependen del trabajo manual de comités que, sin herramientas tecnológicas adecuadas, no siempre logran un análisis sistemático de todo el plan.

Este trabajo propone un flujo de procesamiento (pipeline) de PLN para la organización temática automática de programas de estudio de ingeniería en español. Se comparan técnicas de modelado de tópicos (LDA y NMF) y de agrupamiento (K-Means y jerárquico aglomerativo) sobre un corpus de 51 programas de estudio de la carrera de Ingeniería en Gestión Empresarial. Se evalúan métricas intrínsecas de coherencia y calidad de agrupamiento, y se analiza el efecto de la reducción dimensional mediante SVD sobre el rendimiento del agrupamiento. El objetivo es determinar qué combinación de técnicas produce la organización temática más coherente e interpretable de documentos curriculares.

II. Trabajo relacionado

La aplicación de técnicas de PLN a documentos educativos ha sido abordada desde diferentes perspectivas, aunque con escasa atención a documentos curriculares institucionales. Goštautaitė y Sakalauskas [6] desarrollaron un sistema de clasificación multietiqueta para predecir estilos de aprendizaje a partir de textos educativos, demostrando la viabilidad de aplicar PLN a textos educativos estructurados, aunque sobre materiales en inglés. Yamauchi [14] propuso un método de etiquetado automático de ejercicios educativos en PDF mediante representaciones vectoriales (embeddings), relevante por operar sobre PDFs con estructura variable, aunque sobre ejercicios individuales. Pawar et al. [16] aplicaron técnicas de PLN para la alineación curricular en programas universitarios, evidenciando que la representación textual adecuada del contenido curricular permite identificar redundancias y vacíos. Sekiya et al. [17] exploraron agrupamiento de syllabi universitarios mediante topic models, aunque sobre corpus monodisciplinar.

En modelado de tópicos, Jelodar et al. [4] revisaron LDA y sus variantes para corpus académicos, aunque las aplicaciones operan predominantemente sobre artículos científicos. Qiang et al. [7] concluyeron que NMF produce tópicos más coherentes en corpus con documentos breves, hallazgo relevante dado que los programas de estudio, tras filtrar secciones administrativas, son documentos relativamente cortos. Ferreira-Mello et al. [2] documentaron aplicaciones de minería de texto en educación con escasa atención a documentos curriculares institucionales en español. Salas-Pilco y Xiao [11] identificaron la necesidad de herramientas de IA para educación superior latinoamericana que operen en español y se adapten a estructuras institucionales locales. No se encontraron trabajos que apliquen modelado de tópicos o agrupamiento a programas de estudio completos de educación superior en español.

La contribución del presente trabajo se diferencia en tres aspectos: (1) opera sobre documentos curriculares institucionales completos; (2) procesa textos en español con vocabulario técnico-académico del dominio; (3) compara sistemáticamente múltiples técnicas con métricas intrínsecas estandarizadas, incluyendo el efecto de la reducción dimensional sobre corpus pequeños y un análisis de robustez por variabilidad de inicialización aleatoria y por comparación con representaciones contextuales preentrenadas.

III. Metodología

El enfoque metodológico se organiza en cuatro fases secuenciales. La primera extrae texto de los PDFs originales, conserva las secciones pedagógicamente relevantes y aplica preprocesamiento léxico (tokenización, eliminación de palabras vacías, reducción a la raíz mediante stemming, detección de n-gramas). La segunda construye representaciones vectoriales TF-IDF y aplica modelado de tópicos (LDA, NMF) y agrupamiento (K-Means, Ward jerárquico) con barridos

ISSN 2395-8618

sistemáticos del número de grupos K. La tercera evalúa el efecto de la reducción dimensional mediante SVD sobre la calidad del agrupamiento. La cuarta cuantifica relaciones entre asignaturas mediante similitud coseno. La Figura 1 ilustra el flujo completo. La Tabla 2 detalla las configuraciones técnicas de cada fase.

A. Corpus

El corpus está compuesto por 51 programas de estudio de la carrera de Ingeniería en Gestión Empresarial (IGE) de un instituto tecnológico del sistema nacional de México. Los 51 documentos incluyen las 42 asignaturas del retículo oficial, más 9 asignaturas optativas y de especialidad ofertadas en el campus. Cada programa corresponde a una asignatura y está disponible en formato PDF con estructura estandarizada que incluye: caracterización de la asignatura, intención didáctica, competencias específicas, competencias previas, y temario con contenidos por unidad.

Los documentos cubren el plan de estudios completo, incluyendo asignaturas de ciencias básicas (cálculo, álgebra, física, química), ciencias económico-administrativas (economía, finanzas, contabilidad, mercadotecnia), formación social y humanista (ética, desarrollo humano, desarrollo sustentable), marco jurídico (legislación laboral, propiedad intelectual), metodología de investigación, y asignaturas de especialidad en gestión empresarial, producción y tecnologías de información. La Tabla 1 resume las características del corpus.

TABLA I
DESCRIPCIÓN DEL CORPUS DE PROGRAMAS DE ESTUDIO.

Característica	Valor
Documentos (asignaturas)	51
Tokens totales (post-procesamiento)	22,824
Tokens promedio por documento	448
Tokens mínimo / máximo	237 / 1,112
Vocabulario (raíces únicas)	3,083
Atributos TF-IDF (después de filtrado)	1,653

B. Flujo de procesamiento

El flujo se organiza en cuatro fases (Tabla 2). La extracción de texto conservó secciones con contenido pedagógico: caracterización, intención didáctica, competencias específicas y previas, y temario. Se eliminaron datos generales, participantes en diseño curricular, actividades de aprendizaje, prácticas, criterios de evaluación y fuentes de información, así como encabezados institucionales repetitivos.

Sobre el conjunto estándar de palabras vacías del español se añadió un listado de 80 términos específicos del dominio académico-institucional cuya alta frecuencia transversal a todas las asignaturas no aporta señal discriminativa para la organización temática. Estos términos incluyen vocabulario operativo del programa (asignatura, unidad, tema, subtema, temario, competencia, contenido), agentes y roles institucionales (alumno, estudiante, profesor, docente, academia, instituto, campus), elementos administrativo-pedagógicos (actividad, aprendizaje, evaluación, criterio, hora, crédito, semestre, perfil, egreso), referencias temporales (meses del año, fechas) y nombres institucionales recurrentes (TecNM, Tecnológico Nacional de México, Secretaría Académica).

TABLA II
FLUJO DE PLN PARA ORGANIZACIÓN TEMÁTICA DE PROGRAMAS DE ESTUDIO.

Fase	Etapa	Configuración
1	Extracción	pdfplumber sobre PDFs; conservar secciones pedagógicas (presentación, competencias, temario); eliminar secciones administrativas
1	Tokenización	Minúsculas, eliminar no-alfabéticos, palabras vacías estándar + 80 académicas del dominio
1	Stemming	Snowball español [9]; eliminar tokens <3 caracteres
1	N-gramas	Bigramas/trigramas (min_count=3, threshold=10)
2	TF-IDF	max_df=0.85, min_df=2, max_features=2000
2	Modelado de tópicos	LDA [3] sobre BoW; NMF [8] sobre TF-IDF; $K \in \{3, 5, 7, 8, 10, 12, 15\}$; métricas C_v [10] y C_{npmi}
3	SVD	$d \in \{10, 20, 30, 50\}$ componentes, normalización L2 [1]
3	Agrupamiento	K-Means (n_init=20) y Ward jerárquico; $K \in \{3, 5, 7, 8, 10, 12\}$; Silhouette, Davies-Bouldin, Calinski-Harabasz
4	Similitud	Matriz coseno N×N; identificación de pares extremos

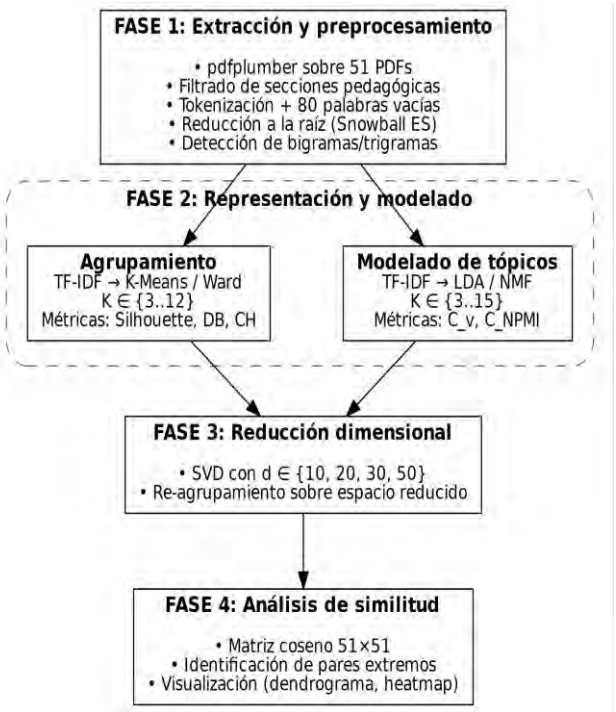

Fig. 1. Flujo de procesamiento propuesto para la organización temática automática de programas de estudio.

ISSN 2395-8618

TABLA III
RESULTADOS DEL BARRIDO DE K (K=3–12). MODELADO DE TÓPICOS EVALUADO CON C_v; AGRUPAMIENTO SOBRE TF-IDF CRUDO EVALUADO CON SILHOUETTE (SIL.) Y DAVIES-BOULDIN (DB).

Mod.	K	C_v	C_{npmi}	Ajuste†	Mod.	K	Sil. / DB / CH
LDA	3	.326	−.031	−6.80	KM	3	.019 / 3.80 / 1.69
LDA	5	.293	−.058	−6.82	KM	5	.022 / 3.51 / 1.61
LDA	7	.300	−.096	−6.82	KM	7	.029 / 3.15 / 1.63
LDA	8	.288	−.121	−6.84	KM	8	.031 / 2.91 / 1.62
LDA	10	.291	−.110	−6.83	KM	10	.039 / 2.60 / 1.62
LDA	12	.307	−.122	−6.85	KM	12	.041 / 2.38 / 1.60
NMF	3	.513	.012	6.46	HW	3	.020 / 4.12 / 1.84
NMF	5	.518	−.043	6.22	HW	5	.031 / 3.35 / 1.81
NMF	7	.611	−.019	6.00	HW	7	.040 / 2.91 / 1.76
NMF	8	.630	−.010	5.89	HW	8	.044 / 2.80 / 1.75
NMF	10	.622	−.029	5.69	HW	10	.048 / 2.48 / 1.72
NMF	12	.627	−.041	5.48	HW	12	.053 / 2.21 / 1.69

KM = K-Means; HW = Jerárquico Ward. †Ajuste: log-perplejidad para LDA (menor = mejor); error de reconstrucción para NMF (menor = mejor).

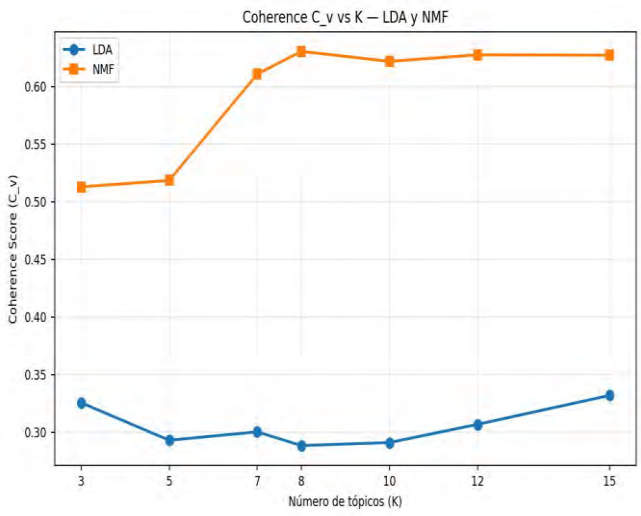

Fig. 2. Coherencia C_v en función del número de tópicos K para LDA y NMF. NMF supera a LDA en todo el rango, con máximo en K=8.

El criterio de selección consistió en identificar términos con frecuencia superior a 30 ocurrencias presentes en más del 70 % de los documentos: la inclusión de estos términos en el vocabulario activo introduce un sesgo de fondo común que enmascara las diferencias temáticas reales entre asignaturas.

La detección de n-gramas con gensim capturó expresiones como "toma_decisión", "gestión_empresarial" o "inteligencia_artificial". Todos los modelos se ejecutaron con random_state=42 para reproducibilidad inicial.

TABLA IV
MÉTRICAS DE AGRUPAMIENTO SOBRE TF-IDF CRUDO VS. TF-IDF+SVD (D=10, 30.3 % VARIANZA EXPLICADA). LA REDUCCIÓN DIMENSIONAL MEJORA TODAS LAS MÉTRICAS POR UN ORDEN DE MAGNITUD.

Método	Representación	K	Silhouette / DB / CH
K-Means	TF-IDF crudo	12	0.041 / 2.383 / 1.60
K-Means	TF-IDF+SVD(10)	10	0.418 / 0.787 / 14.84
Jerárquico	TF-IDF crudo	12	0.053 / 2.207 / 1.69
Jerárquico	TF-IDF+SVD(10)	10	0.401 / 0.847 / 13.86

IV. RESULTADOS

C. Modelado de tópicos: LDA vs. NMF

La Figura 2 muestra la coherencia C_v en función del número de tópicos K para ambos modelos. NMF supera consistentemente a LDA en todo el rango evaluado. LDA alcanza un máximo de C_v=0.33 con K=15, mientras que NMF obtiene C_v=0.63 con K=8.

La Tabla III presenta los resultados completos del barrido de K para los cuatro modelos. En modelado de tópicos, NMF domina a LDA en C_v para todas las configuraciones. En agrupamiento, los valores de Silhouette sobre TF-IDF crudo son cercanos a cero independientemente de K, lo cual motiva la reducción dimensional analizada en la siguiente sección.

Este resultado es consistente con la literatura sobre corpus pequeños: LDA, como modelo generativo bayesiano, requiere mayor volumen de datos para estimar distribuciones confiables [4], mientras que NMF, al operar directamente sobre la matriz TF-IDF mediante factorización algebraica, aprovecha mejor la información disponible en colecciones reducidas.

D. Efecto de la reducción dimensional en agrupamiento

El agrupamiento sobre la matriz TF-IDF cruda (51 × 1,653) produjo valores de Silhouette cercanos a cero (Tabla 3), indicando ausencia de estructura de agrupamiento detectable en el espacio original. El ratio atributos/documentos (≈ 32:1) y el vocabulario compartido entre asignaturas de una misma carrera explican este resultado.

La aplicación de SVD como paso previo al agrupamiento transformó los resultados. La Figura 3 muestra el Silhouette Score comparando TF-IDF crudo contra distintas dimensiones SVD. La Tabla 4 resume la mejora con la mejor configuración.

La mejora es de un orden de magnitud en Silhouette (de 0.04 a 0.42), elevando el resultado de "sin estructura detectable" a "estructura razonable". El hecho de que d=10 (30.3 % de varianza explicada) supere a d=20, 30 y 50 se explica por la naturaleza de la descomposición: las primeras componentes singulares capturan la varianza inter-grupo, es decir, las diferencias entre áreas temáticas (matemáticas vs. derecho vs. mercadotecnia). Las componentes adicionales capturan

ISSN 2395-8618

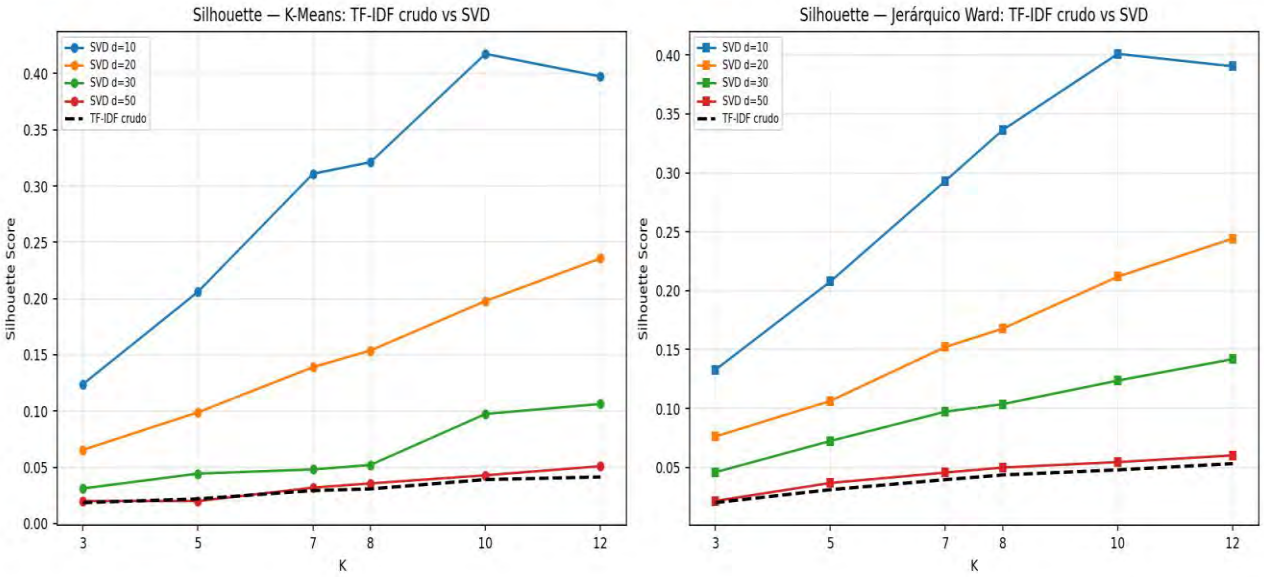

Fig. 3. Silhouette Score vs. K para K-Means (izquierda) y jerárquico (derecha), comparando TF-IDF crudo (línea punteada) contra TF-IDF con SVD a distintas dimensiones.

TABLA V
TÓPICOS NMF (K = 8): ETIQUETA SEMÁNTICA, PALABRAS CLAVE Y ASIGNATURAS CON MAYOR CARGA (PESO NORMALIZADO).

T	Etiqueta	Palabras clave	Asignaturas principales
0	Producción y costos	producción, costo, empresa, administración, control	Costos Emp. (1.00), Gestión Prod. II (.93), Presupuestación (.93)
1	Desarrollo humano y ética	humano, social, ética, persona, grupo	Desarrollo Humano (1.00), Dinámica Social (.97), Hab. Blandas (.93)
2	Mercadotecnia y negocios digitales	mercadotecnia, modelo_negocio, datos, inteligencia_negocios	Innov. Tecnológica (1.00), Merc. Electrónica (1.00), Mercadotecnia (.75)
3	Metodología de investigación	investigación, protocolo, taller, documental	Taller Inv. I (.98), Taller Inv. II (.94), Fund. Investigación (.92)
4	Estadística y calidad	varianza, calidad, muestreo, estadística, prueba	Est. Inferencial I (1.00), Est. Inferencial II (.94), Calidad (.71)
5	Matemáticas y ciencias básicas	función, cálculo_integral, definición, matriz	Álgebra Lineal (1.00), Cálculo Integral (1.00), Cálculo Diferencial (.96)
6	Economía y sustentabilidad	economía, macroeconomía, equilibrio, mercado	Entorno Macro. (1.00), Economía Emp. (.93), Com. Internacional (.59)
7	Marco jurídico-laboral	trabajo, derecho, laboral, contrato, seguridad_social	Legislación Lab. (1.00), Estrat. Contables (.91), Marco Legal (.82)

progresivamente varianza intra-grupo: diferencias estilísticas, variaciones en extensión o formulaciones particulares de cada programa que no aportan a la discriminación temática y actúan como ruido para el algoritmo de agrupamiento.

E. Análisis de tópicos NMF (K = 8)

La Tabla 5 presenta los 8 tópicos generados por NMF con su interpretación cualitativa. Cada tópico se etiquetó a partir de las palabras con mayor peso y se reportan las asignaturas con mayor carga. Las etiquetas fueron asignadas por un investigador con conocimiento del plan de estudios; una validación con múltiples evaluadores se reporta en un trabajo complementario [18].

Los 8 tópicos cubren las principales áreas formativas del plan de estudios de manera coherente. La distribución de tópico dominante por asignatura es balanceada: T0 (12 asignaturas), T2 (9), T1 (7), T7 (6), T5 (5), T6 (5), T4 (4) y T3 (3). No se observan tópicos degenerados.

La Figura 4 muestra la distribución de tópicos por asignatura. La mayoría presenta un tópico dominante claramente definido, pero varias exhiben cargas distribuidas en múltiples tópicos, reflejando el carácter interdisciplinario de ciertas materias. Por ejemplo, Plan de Negocios combina cargas en T0 (producción/costos), T2 (mercadotecnia/negocios) y T6 (economía); Desarrollo Sustentable combina T1 (ética/social) con T6 (economía/sustentabilidad). Esta capacidad de detectar asignaturas multitemáticas es una ventaja del modelado de tópicos frente al agrupamiento duro.

F. Agrupamiento jerárquico y similitud entre asignaturas

La Figura 5 presenta el dendrograma del agrupamiento jerárquico (Ward) sobre TF-IDF+SVD(10) con corte en K = 10. Las agrupaciones resultantes presentan alta correspondencia con la estructura disciplinar esperada: matemáticas (3 asignaturas), investigación (3), estadística y calidad (4), economía (3), finanzas y contabilidad (6), marco jurídico-laboral (5), desarrollo humano y social (6), gestión y administración (8), mercadotecnia y tecnologías de información (8), y producción e ingeniería industrial (5). Los tamaños oscilan entre 3 y 8 documentos (ratio máximo/mínimo de 2.7), sin agrupamientos degenerados.

ISSN 2395-8618

Fig. 4. Distribución de tópicos NMF (K = 8) por asignatura. Intensidad proporcional al peso normalizado. Asignaturas interdisciplinarias presentan cargas en múltiples tópicos.

La convergencia entre modelado de tópicos y agrupamiento es notable: los 8 tópicos NMF y los 10 agrupamientos jerárquicos coinciden en las agrupaciones nucleares (matemáticas, investigación, estadística, jurídico), con discrepancias concentradas en asignaturas de frontera entre áreas disciplinares.

La Figura 6 muestra la matriz de similitud coseno entre las 51 asignaturas, ordenada por agrupamiento jerárquico. Los bloques diagonales oscuros confirman la cohesión intra-agrupamiento, mientras que las regiones claras entre bloques evidencian la separación inter-agrupamiento.

Se observan conexiones moderadas entre los bloques de economía y finanzas, así como entre gestión y producción, lo cual refleja relaciones interdisciplinarias reales del plan de estudios.

El análisis de pares extremos validó la coherencia del flujo: las similitudes máximas se observaron entre asignaturas secuenciales (Taller de Investigación I–II, 0.54) y dentro de áreas afines (Legislación Laboral–Estrategias Contables, 0.39); las mínimas (≈0.02) cruzan áreas disciplinares opuestas.

G.Análisis de robustez y comparación con representaciones contextuales

Para evaluar la robustez ante variabilidad estocástica, se replicó la configuración óptima (NMF K=8; agrupamiento K=10 sobre SVD de 10 componentes) con 10 semillas aleatorias. La coherencia C_v se mantuvo en 0.606 ± 0.021 (CV=3.4 %), el Silhouette de K-Means en 0.401 ± 0.010 (CV=2.5 %) y el de Ward en 0.383 ± 0.012 (CV=3.1 %). La estabilidad estructural por Adjusted Rand Index (ARI) entre las 45 comparaciones por pares fue 0.918 ± 0.049 para K-Means y 0.783 ± 0.080 para Ward.

Los coeficientes de variación bajo 5 % y los ARI altos confirman que los hallazgos no dependen de la inicialización aleatoria.

Como representación alternativa, se generaron embeddings con paraphrase-multilingual-MiniLM-L12-v2 (384 dimensiones), reducidos también con SVD a 10 componentes para comparación equitativa. K-Means con K=10 alcanzó Silhouette de 0.417 sobre TF-IDF+SVD frente a 0.223 sobre

ISSN 2395-8618

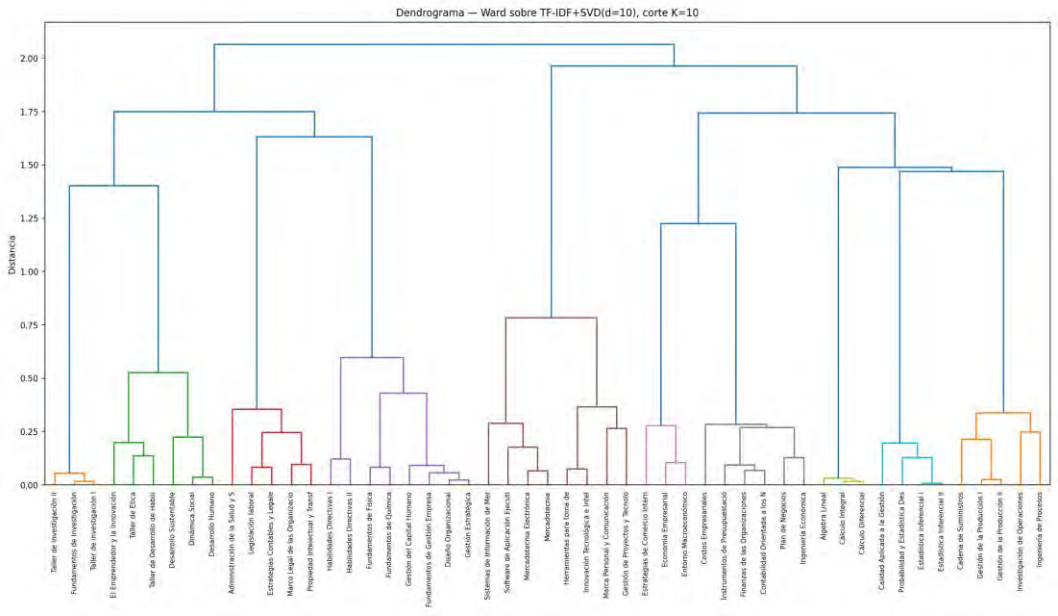

Fig. 5. Dendrograma del agrupamiento jerárquico (Ward) sobre TF-IDF+SVD(d=10) con corte en K = 10.

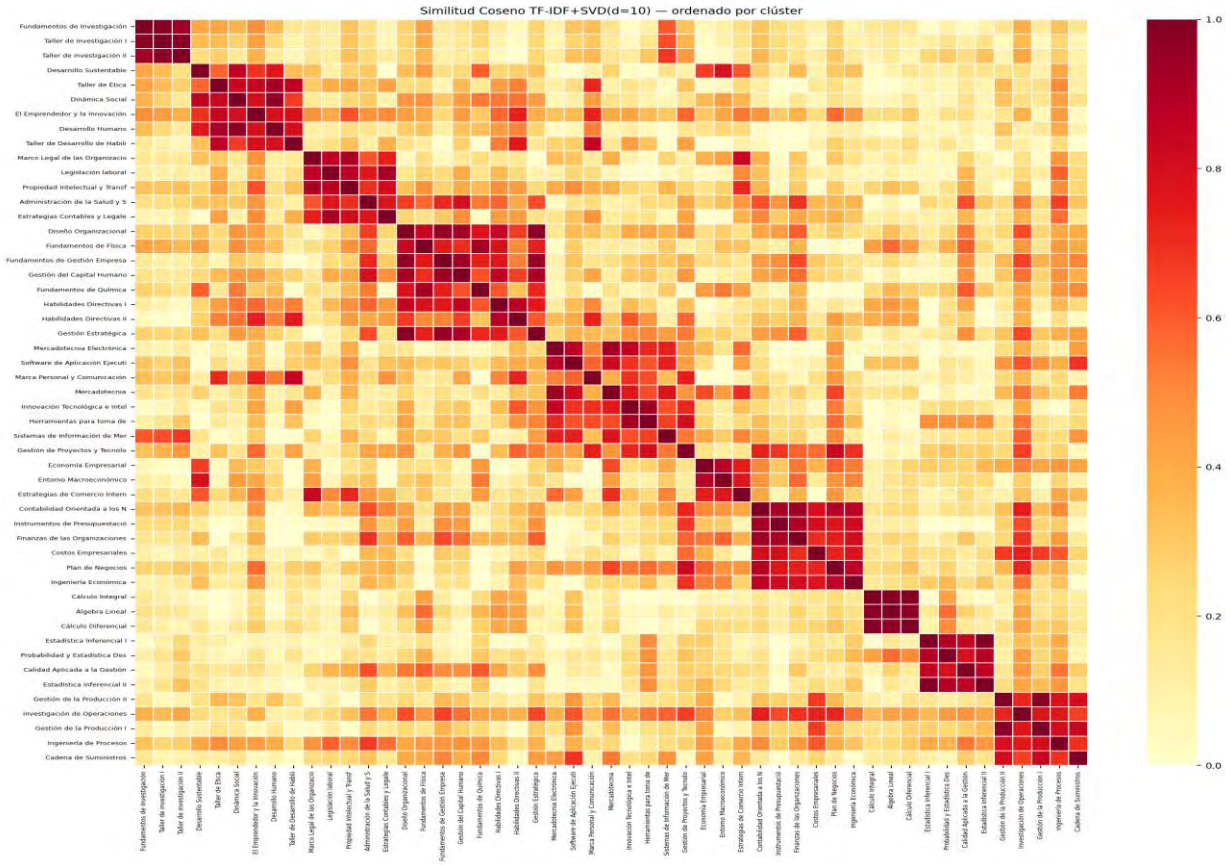

Fig. 6. Matriz de similitud coseno entre asignaturas (TF-IDF+SVD), ordenada por agrupamiento jerárquico. Los bloques diagonales corresponden a agrupamientos con alta cohesión interna.

embeddings+SVD; el ARI entre ambas particiones fue 0.049 (cercano al azar). Las dos representaciones detectan estructuras distintas: TF-IDF aprovecha la señal disciplinar del vocabulario técnico, mientras que los embeddings preentrenados sin adaptación de dominio capturan similitudes semánticas más genéricas que diluyen las fronteras disciplinares.

ISSN 2395-8618

Para organización temática curricular, donde la separación disciplinar es el criterio principal, TF-IDF+SVD es la representación adecuada.

V. Discusión

NMF como modelo preferente para modelado de tópicos curricular. NMF superó a LDA de manera consistente, confirmando hallazgos de Qiang et al. [7] sobre su superioridad en corpus con documentos cortos. Para planes de estudio individuales (40–55 documentos), NMF es la opción recomendada. LDA podría ser más competitivo al escalar a los 41 planes del sistema (≈1,800 documentos).

La reducción dimensional en corpus académicos pequeños. SVD con 10 componentes (30 % de varianza) produjo la mejor separación al capturar las diferencias temáticas gruesas entre áreas disciplinares, eliminando la varianza intra-grupo (variaciones estilísticas y de extensión) que actúa como ruido. Para corpus académicos pequeños con vocabulario disciplinar compartido como el aquí estudiado, la reducción dimensional resulta indispensable; la generalización a corpus de mayor tamaño o composición distinta requiere validación empírica adicional.

Valor complementario de modelado de tópicos y agrupamiento. NMF permite pertenencia múltiple, identificando asignaturas interdisciplinarias; el agrupamiento genera particiones discretas útiles para organización administrativa. Su concordancia refuerza la validez: cuando técnicas distintas convergen, la estructura detectada es probablemente real y no un artefacto del método.

Aplicaciones prácticas. El flujo aquí presentado puede operacionalizarse como herramienta de apoyo a procesos de revisión curricular en academias del sistema TecNM. Las salidas del agrupamiento jerárquico permiten visualizar la organización temática del plan a comités académicos sin formación técnica especializada, identificar solapamientos potenciales entre asignaturas (pares con similitud coseno alta dentro de distintas áreas) y detectar asignaturas con perfil multitemático que podrían requerir reformulación o reasignación. La distribución de tópicos NMF complementa este análisis ofreciendo una mirada cuantitativa sobre la composición temática efectiva del plan, útil tanto para reportes institucionales como para análisis comparativos entre planes de carrera o entre cohortes de actualización curricular.

Limitaciones. El corpus (51 documentos) constituye un plan completo pero podría no generalizar a carreras con estructura diferente. La evaluación intrínseca aquí presentada se complementa con la validación extrínseca multi-evaluador reportada en un trabajo paralelo [18]. El agrupamiento de "gestión y administración" agrupa asignaturas diversas (incluyendo Física y Química) porque sus programas emplean vocabulario orientado a aplicaciones que, tras reducción a la raíz, se confunde con el léxico administrativo; representaciones basadas en transformadores contextuales con adaptación de dominio podrían resolver esta limitación, dirección que la comparación realizada con representaciones genéricas preentrenadas (Sección 4.5) sugiere explorar como trabajo futuro.

VI. Conclusiones y trabajo futuro

Este trabajo presentó un flujo de PLN para la organización temática automática de programas de estudio de ingeniería en español. Los resultados principales son:

NMF con K = 8 tópicos produjeron la organización más coherente (C_v=0.63), superando a LDA (C_v=0.33) en todo el rango. Los 8 tópicos son interpretables y corresponden a áreas formativas reconocibles del plan de estudios.

La reducción dimensional mediante SVD es indispensable para el agrupamiento de documentos curriculares en corpus pequeños con vocabulario disciplinar compartido: SVD con 10 componentes mejoró el Silhouette de 0.04 a 0.42 y el Davies-Bouldin de 2.38 a 0.79, al aislar la varianza inter-grupo que discrimina entre áreas temáticas.

El agrupamiento jerárquico (Ward) sobre TF-IDF+SVD produjo 10 agrupaciones disciplinariamente coherentes con distribución balanceada (3–8 documentos por agrupamiento), con alta correspondencia con los tópicos NMF.

El análisis de similitud coseno identificó relaciones esperadas entre asignaturas secuenciales y áreas afines, validando la capacidad del flujo para capturar la estructura curricular.

Análisis de robustez (CV<5 % entre semillas, ARI=0.92 entre particiones) y comparación con representaciones contextuales preentrenadas (TF-IDF+SVD supera a embeddings+SVD: 0.42 vs 0.22 en Silhouette) confirman la solidez del enfoque para la tarea específica de organización temática curricular.

Como trabajo futuro se plantea: (a) expansión a los 41 planes de ingeniería del sistema nacional de institutos tecnológicos (≈1,800 programas); (b) exploración de embeddings contextuales con adaptación de dominio (fine-tuning) sobre vocabulario académico mexicano, dado que las representaciones preentrenadas genéricas mostraron capturar estructuras distintas a las disciplinares; (c) desarrollo de una herramienta interactiva para visualización de estructura temática curricular orientada a comités académicos.

Referencias

[1] Aggarwal, "Mining Text Data," in *Data Mining*, Springer, 2015, doi: 10.1007/978-3-319-14142-8_13.

[2] R. Ferreira-Mello, M. André, A. Pinheiro, E. Costa, and C. Romero, "Text mining in education," *WIREs Data Min. Knowl. Discov.*, vol. 9, no. 6, p. e1332, 2019, doi: 10.1002/widm.1332.

[3] M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[4] H. Jelodar, Y. Wang, C. Yuan, X. Feng, X. Jiang, Y. Li, and L. Zhao, "Latent Dirichlet Allocation (LDA) and topic modeling: models, applications, a survey," *Multimedia Tools Appl.*, vol. 78, pp. 15169–15211, 2019, doi: 10.1007/s11042-018-6894-4.

[5] F. J. García-Peñalvo, "La percepción de la Inteligencia Artificial en contextos educativos tras el lanzamiento de ChatGPT: disrupción o pánico," *Educ. Knowl. Soc.*, vol. 24, p. e31279, 2023, doi: 10.14201/eks.31279.

[6] Goštautaitė and L. Sakalauskas, "Multi-label classification and explanation methods for students' learning style

ISSN 2395-8618

prediction and interpretation," *Appl. Sci.*, vol. 12, no. 11, p. 5396, 2022, doi: 10.3390/app12115396.

[7] J. Qiang, Z. Qian, Y. Li, Y. Yuan, and X. Wu, "Short text topic modeling techniques, applications, and performance," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 3, pp. 1427–1445, 2022, doi: 10.1109/TKDE.2020.2992485.

[8] D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999, doi: 10.1038/44565.

[9] M. F. Porter, "An algorithm for suffix stripping," *Program*, vol. 14, no. 3, pp. 130–137, 1980, doi: 10.1108/eb046814.

[10] M. Röder, A. Both, and A. Hinneburg, "Exploring the space of topic coherence measures," in *Proc. 8th ACM Int. Conf. Web Search Data Min. (WSDM)*, 2015, pp. 399–408, doi: 10.1145/2684822.2685324.

[11] S. Z. Salas-Pilco and K. Xiao, "Artificial intelligence applications in Latin American higher education: a systematic review," *Int. J. Educ. Technol. Higher Educ.*, vol. 17, p. 21, 2020, doi: 10.1186/s41239-020-00200-0.

[12] Tecnológico Nacional de México, *Anuario estadístico del Tecnológico Nacional de México 2022–2023*, TecNM, 2023.

[13] S. Wang, C. Christensen, T. Xu, S. Cui, R. Han, C. Zhuo, and J. Xie, "Artificial intelligence in education: a systematic literature review," *J. Res. Technol. Educ.*, 2024, doi: 10.1080/15391523.2024.2314488.

[14] T. Yamauchi, "Automated labeling of educational PDF exercises using word embedding and classification methods," *Smart Learn. Environ.*, vol. 10, p. 7, 2023, doi: 10.1186/s40561-023-00271-9.

[15] O. Zawacki-Richter, V. I. Marín, M. Bond, and F. Gouverneur, "Systematic review of research on artificial intelligence applications in higher education," *Int. J. Educ. Technol. Higher Educ.*, vol. 16, p. 39, 2019, doi: 10.1186/s41239-019-0171-0.

[16] Pawar and V. Mago, "Calculating the similarity between words and sentences using a lexical database and corpus statistics," arXiv:1802.05667, 2018.

[17] T. Sekiya, Y. Matsuda, and K. Yamaguchi, "Curriculum analysis of CS departments based on CS2013 by simplified, supervised LDA," in *Proc. 5th Int. Conf. Learn. Anal. Knowl. (LAK)*, 2015, pp. 330–339.

[18] I. Abraján Barraza, A. S. de la Serna Tuya, K. L. Carvajal Estrada, and I. A. Palazuelos Gálvez, "Validation of an NLP model for thematic organization of study programs: teacher diagnosis, gold standard and extrinsic evaluation," in *Proc. 6th Congr. Int. Tecnol. Cienc. Apl. (CITCA)*, 2026 [in press].