



UNIVERSIDAD AUTÓNOMA DE SAN
LUIS POTOSÍ
FACULTAD DE ECONOMÍA



Maestría en Economía Matemática

Análisis de las Propuestas Legislativas en el Congreso de San Luis Potosí (2016-2022)

TESIS

Para obtener el grado de:
Maestro en Economía Matemática

Presenta:

Ortiz Cuenca Cristhian Alonso

Tutor:

Dr. Jorge Zazueta Gutierrez

Sinodales:

Dr. Leonardo David Tenorio Martínez

Dr. Leobardo Pedro Plata Pérez

Agosto del 2024, San Luis Potosí, S.L.P., México

Folio Num. 045



"Análisis de las Propuestas Legislativas en el Congreso de San Luís Potosí (2016-2022)"

© 2024 by Cristhian Alonso Ortiz Cuenca is licensed under [Creative Commons](https://creativecommons.org/licenses/by-nc-nd/4.0/)

[Attribution-NonCommercial-NoDerivatives 4.0 International](https://creativecommons.org/licenses/by-nc-nd/4.0/)

Agradecimientos

Quiero expresar mi más profundo agradecimiento a todas las personas que hicieron posible la realización de esta tesis.

A mi familia, por ser el pilar fundamental en todos mis proyectos y por su incondicional apoyo.

A mi director de tesis, Jorge Zazueta, por su invaluable tiempo y paciencia, y por guiarme con dedicación durante todo el proceso de elaboración de esta tesis.

A mis sinodales, Leonardo y Leobardo, por sus correcciones y valiosas ideas que contribuyeron significativamente a mejorar este trabajo.

A mis profesores de la maestría, por el tiempo dedicado a enseñarnos y compartir su conocimiento con nosotros.

A mis compañeros, por todas esas horas de colaboración y esfuerzo conjunto en la realización de tareas y proyectos.

Finalmente, quiero agradecer a CONAHCYT por financiar mis estudios en México, haciendo posible este logro académico.

A todos ustedes, muchas gracias.

Cristhian Ortiz

Resumen

Esta tesis ofrece un análisis exhaustivo de las propuestas legislativas presentadas en el Congreso de San Luis Potosí durante las legislaturas LXI, LXII y LXIII (2016-2022), utilizando técnicas avanzadas de procesamiento de lenguaje natural y análisis de tópicos. A través del modelo Latent Dirichlet Allocation (LDA), se identificaron y categorizaron 17 tópicos en más de 3,600 documentos legislativos. El estudio revela la predominancia de temas administrativos y burocráticos sobre los tópicos de interés social, con una proporción que disminuye ligeramente del 40.86% en 2016 al 36.75% en 2022, indicando un aumento gradual en el enfoque hacia temas sociales.

Los resultados destacan la evolución de las prioridades legislativas, evidenciando una constante en temas de justicia social y desarrollo inclusivo, y una variabilidad significativa en la relevancia de la auditoría. El impacto de la pandemia de COVID-19 se refleja en el aumento temporal de las propuestas relacionadas con la salud comunitaria. El análisis de aglomeración y la visualización de similitudes temáticas entre partidos políticos muestran patrones de afinidad, como la similitud entre el gobernador y PVEM, y diferencias significativas entre partidos como PR y PRIVADA, y RSP y PR.

Se observan cambios notables en tópicos específicos: los temas relacionados con la salud comunitaria y la protección a víctimas han ganado importancia, mientras que los tópicos sanitarios relacionados con COVID-19 han disminuido su relevancia post-pandemia. Además, los tópicos de leyes de tránsito y transporte público han fluctuado, posiblemente debido a cambios en la administración y políticas de movilidad.

Este estudio no solo proporciona una comprensión integral de las dinámicas legislativas en San Luis Potosí, sino que también ofrece una base empírica sólida para futuras investigaciones y la formulación de políticas más informadas. La investigación demuestra los beneficios de un análisis cuantitativo en un ámbito tradicionalmente cualitativo, proporcionando una manera eficiente de medir la actividad política y legislativa.

Palabras Clave: Propuestas Legislativas, Congreso de San Luis Potosí, Latent Dirichlet Allocation (LDA), Clustering de Partidos, Hierarchical Clustering, Distribución de Propuestas, Clustering de Partidos.

Tabla de contenidos

1	Introducción, Objetivo y Revisión de la Literatura	6
1.1	Introducción	6
1.2	Objetivo	7
1.3	Revisión de la Literatura	7
2	Notación Básica y Terminología.	9
2.1	Distribución de Dirichlet	10
3	LDA (Latent Dirichlet Allocation)	12
3.1	Presentación del modelo LDA	12
3.2	Proceso Generativo	15
4	Estimación de LDA mediante Muestreo de Gibbs	19
4.0.1	Estimación de Parámetros (ϕ y θ).	20
5	LDA aplicado a las Propuestas Legislativas del Congreso de San Luis Potosí	21
5.1	Descripción del Proceso	21
6	Análisis de Resultados	24
6.1	Base de Datos	24
6.2	Análisis de Tópicos Latentes en Datos del Congreso de S.L.P.	26
6.3	Análisis Comparativo de Rankings y Proporciones	32
6.4	Cambio de Ranking por Tópico y por Legislatura	34
6.5	Prioridades Legislativas de los Partidos Dominantes	38
6.6	Análisis Jerárquico de Aglomeración	42
6.6.1	Dendograma	42
6.6.2	Análisis Jerárquico y Visualización de Similitudes entre Proponentes	42
6.7	Aportes de la Tesis	46
7	Apéndices	48
7.1	Partidos predominantes	48
7.2	Métricas para Determinar el Número de Tópicos en LDA	50
7.3	Descripción del Proceso LDA	51
7.4	Cálculo del Cambio de Ranking de los Tópicos Legislativos	54
7.4.1	Cálculo del Cambio Total en el Ranking	54
7.5	Ejemplo computacional	57
8	Conclusiones	65
	Bibliografía	67

1 Introducción, Objetivo y Revisión de la Literatura

1.1 Introducción

En el contexto del Congreso del Estado de San Luis Potosí, el análisis de propuestas legislativas es fundamental para comprender cómo han evolucionado las prioridades políticas y legislativas a lo largo del tiempo. Este trabajo emplea técnicas avanzadas para examinar más de 3,600 documentos legislativos de las tres últimas legislaturas (LXI, LXII, LXIII) durante el periodo 2016-2022, con el objetivo de identificar las principales áreas temáticas y sus cambios a lo largo de este tiempo.

El Congreso de San Luis Potosí, establecido en 1824, ha sido un actor clave en la creación y modificación de leyes que afectan a la vida de los ciudadanos. A lo largo de su historia, ha experimentado importantes transformaciones, reflejando reformas políticas y sociales significativas. En las últimas décadas, ha evolucionado de un órgano meramente ratificador a uno más proactivo en la formulación y debate de propuestas legislativas, destacándose en la discusión de temas como el sufragio femenino y otras reformas estructurales Argüelles, Elia (2017).

La cantidad de texto no es difícil de analizar; el texto es difícil de analizar debido al volumen de información. Dado el volumen y la complejidad de la información contenida en las propuestas legislativas, esta investigación utiliza el modelo Latent Dirichlet Allocation (LDA) para organizar y analizar los documentos de manera eficiente. El LDA, un modelo generativo en minería de textos, permite identificar patrones latentes en grandes volúmenes de texto al agrupar palabras que suelen ocurrir juntas en tópicos específicos. Este enfoque facilita la identificación de tendencias y cambios en las prioridades legislativas, y proporciona una base empírica para evaluar cómo eventos históricos y transformaciones sociales, como la pandemia de COVID-19, han influido en las propuestas y en la agenda política.

El análisis revela cómo las prioridades de los legisladores han cambiado en respuesta a factores socioeconómicos y políticos, y permite entender las dinámicas detrás de las decisiones políticas y legislativas en San Luis Potosí. Esta investigación no solo contribuye a una comprensión más profunda de la evolución legislativa en la región, sino que también ofrece una herramienta valiosa para evaluar y contextualizar la actividad política en un periodo de cambio y desafío.

1.2 Objetivo

Analizar las propuestas presentadas por congresistas, ciudadanos, empresas públicas y privadas durante las tres últimas legislaturas del Congreso de San Luis Potosí (LXI, LXII, LXIII), (2016-2022) con el fin de identificar la evolución de las propuestas y prioridades de las diferentes legislaturas.

Preguntas de investigación:

- ¿Cuáles son los tópicos propuestos al congreso?
- ¿Cómo cambia la distribución de tópicos en función de las legislaturas y partido político?
- ¿Cómo cambia la distribución de tópicos en función de algunos eventos históricos (COVID, cambios políticos, etc)?
- ¿Qué tan similares son los partidos políticos en sus propuestas y cómo se alinean estas con su plan de acción?

1.3 Revisión de la Literatura

El análisis de propuestas legislativas utilizando modelos de tópicos no es un concepto nuevo y ha sido aplicado en diversas áreas. Estas metodologías permiten extraer información valiosa de grandes volúmenes de texto, facilitando una comprensión más profunda de las dinámicas políticas y legislativas.

D. M. Blei (2012) destaca que Latent Dirichlet Allocation (LDA) es un modelo generativo que permite descubrir la estructura latente en grandes colecciones de documentos, identificando grupos de palabras que tienden a ocurrir juntas, denominados “tópicos”. Este enfoque es especialmente útil en contextos donde se requiere analizar grandes volúmenes de texto no estructurado para extraer patrones significativos.

En el contexto político, Curran et al. (2018) aplicaron un modelo de LDA para analizar los discursos parlamentarios de Nueva Zelanda, construyendo redes bipartitas que vinculan a los miembros del parlamento con los temas que discuten. Este enfoque permitió identificar cómo cambia la popularidad de los temas a lo largo del tiempo y relacionar las tendencias seguidas por los partidos políticos con eventos sociales, económicos y legislativos específicos. Además, el análisis de la comunidad de las proyecciones de redes bipartitas reveló qué partidos dominan el debate político y hasta qué punto tienden a especializarse en un número pequeño o grande de tópicos.

De manera similar, Luque, Carolina et al. (2021) emplearon un modelo de Asignación Latente de Dirichlet (LDA) para analizar patrones en la investigación del COVID-19, utilizando 4928 resúmenes científicos de PubMed publicados en el primer semestre de 2020. El análisis identificó dos tópicos principales: factores de riesgo y severidad, e impacto en la salud pública, facilitando así una revisión bibliográfica más eficiente. Este enfoque metodológico es altamente

relevante para el análisis de las propuestas legislativas del Congreso de San Luis Potosí, donde se utiliza LDA para identificar y categorizar tópicos recurrentes en las propuestas, optimizando la revisión de la literatura y facilitando la identificación de tópicos emergentes.

I.-C. Chang et al. (2021) aplicaron LDA para clasificar tópicos en revistas de educación ambiental, utilizando técnicas de minería de textos y análisis de conglomerados para interpretar las categorías y características de los artículos publicados entre 2011 y 2020. Este estudio destaca la utilidad de LDA para identificar y categorizar los tópicos principales, facilitando la organización y el acceso a la información relevante. Los resultados demostraron la consistencia y corrección en la clasificación de tópicos, lo que subraya la robustez de LDA en diferentes contextos, incluido el análisis de propuestas legislativas del Congreso de San Luis Potosí, proporcionando una metodología eficiente para la revisión y categorización de grandes volúmenes de texto.

Además, Laver, Benoit, y Garry (2003) presentan una metodología innovadora para extraer posiciones políticas de textos utilizando palabras como datos. Este enfoque cuantifica los textos en lugar de interpretarlos discursivamente, permitiendo replicar las estimaciones de posiciones políticas previamente publicadas de manera eficiente. Aplicaron su técnica para analizar manifiestos de partidos políticos en Gran Bretaña e Irlanda, extendiendo su uso a discursos legislativos.

Odden, Marin, y Rudolph (2021) investigan la evolución de la educación en ciencias a lo largo de los últimos 100 años mediante el uso de técnicas de procesamiento de lenguaje natural, específicamente LDA. Este estudio analiza más de 11,000 artículos publicados en la revista *Science Education* desde 1922 hasta 2019, clasificándolos en 21 tópicos distintos agrupados en tres categorías: tópicos de contenido científico, enfoques en la enseñanza y enfoques centrados en el estudiante. Los autores destacan cómo ciertos tópicos han surgido y disminuido en popularidad a lo largo del tiempo, influenciados por cambios en políticas nacionales y avances en campos adyacentes. Este análisis automatizado ofrece una revisión bibliográfica replicable y una visión macro de las tendencias históricas, que puede ser aplicada en el análisis de la evolución temática de las propuestas legislativas en el Congreso de San Luis Potosí, permitiendo identificar patrones y cambios en las prioridades legislativas a lo largo del tiempo.

La relevancia de estos estudios radica en su capacidad para proporcionar una comprensión integral de las dinámicas políticas a través de metodologías cuantitativas avanzadas. Al combinar técnicas de modelado de tópicos con análisis de redes complejas, se pueden obtener insights valiosos sobre las prioridades y enfoques temáticos en distintos contextos legislativos. Esto no solo facilita la identificación de patrones emergentes, sino que también proporciona una base empírica sólida para futuras investigaciones y la formulación de políticas más informadas y efectivas.

2 Notación Básica y Terminología.

En este trabajo, empleamos terminología relacionada con conjuntos de textos, refiriéndonos a conceptos como “palabras o términos”, “tópicos”, “documentos” y “corpus”.

Definición 1. Palabra: Es la unidad básica de datos discretos, definida como un elemento de un vocabulario indexado por $\{1, \dots, V\}$ (El conjunto de todas las palabras en un corpus se llama vocabulario). Matemáticamente, una palabra w pertenece al vocabulario V , y en el contexto de un documento, las palabras son generadas a partir de las distribuciones de tópicos. Para un documento d , la palabra $w_{d,i}$ en la posición i es generada a partir de una distribución multinomial $\text{Mult}(\beta_{z_{d,i}})$, donde $z_{d,i} = k$ es el tópico asignado a la palabra en esa posición.

Definición 2. Tópico: Es una distribución sobre palabras. Matemáticamente, un tópico β_k es una distribución multinomial sobre el vocabulario V . Si hay K tópicos, entonces β es una matriz $K \times V$, donde V es el número de palabras en el vocabulario. Cada fila de esta matriz corresponde a un tópico, es decir, $\beta_k = (\beta_{k1}, \beta_{k2}, \dots, \beta_{kV})$, donde β_{kv} es la probabilidad de la palabra v en el tópico k .

Definición 3. Documento: Es una distribución sobre tópicos. Matemáticamente, un documento d se representa como una distribución multinomial θ_d sobre los K tópicos. Es decir, $\theta_d = (\theta_{d1}, \theta_{d2}, \dots, \theta_{dK})$, donde θ_{dk} es la probabilidad del tópico k en el documento d .

Definición 4. Corpus: Es una colección de M documentos. Matemáticamente, el corpus se representa como $D = \{\theta_1, \theta_2, \dots, \theta_M\}$ donde cada θ_d es una distribución sobre tópicos para el documento d en el corpus.

Tabla 1: Listado de símbolos

Símbolo	Descripción
K	número especificado de tópicos
k	índice auxiliar sobre los tópicos
V	número de palabras en el vocabulario
v	índice auxiliar sobre el vocabulario
d	índice auxiliar sobre documentos
N_d	longitud del documento (número de palabras)
i	índice auxiliar sobre palabras en un documento
α	vector positivo de dimensión K
β	vector positivo de dimensión V
$\text{Dir}(\alpha)$	una distribución de Dirichlet de K dimensiones
$\text{Dir}(\beta)$	una distribución de Dirichlet de V dimensiones
z	índices de tópicos: $z_{d,i} = k$ significa que la i -ésima palabra en el d -ésimo documento está asignada al tópico k

2.1 Distribución de Dirichlet

La distribución de Dirichlet es una distribución de probabilidad continua ampliamente utilizada en estadística bayesiana y modelos generativos en aprendizaje automático. Se considera la distribución conjugada previa de la distribución multinomial, lo cual facilita la inferencia bayesiana al permitir actualizaciones sencillas de creencias sobre probabilidades categóricas con la incorporación de nuevos datos (Bela A., Frigiyk, Amol, Kapila, y Maya R, Gupta 2010).

La función de densidad de probabilidad (PDF) de la distribución de Dirichlet para un vector $\mathbf{x} = (x_1, x_2, \dots, x_K)$ con parámetros $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_K)$ se define como:

$$p(x | \alpha_1, \dots, \alpha_K) = \frac{\Gamma\left(\sum_{i=1}^K \alpha_i\right)}{\prod_{i=1}^K \Gamma(\alpha_i)} \prod_{i=1}^K x_i^{\alpha_i-1}$$

Aquí, α es un vector positivo de K dimensiones y Γ denota la función Gamma, que es una generalización de la función factorial a valores reales M. E. Roberts, Stewart, y Airoldi (2016). La función Gamma se define como:

$$\Gamma(z) = \int_0^{\infty} t^{z-1} e^{-t} dt$$

Propiedades

1. **Simplex:** Los vectores \mathbf{x} distribuidos según Dirichlet se encuentran en el simplex de dimensiones $K - 1$, asegurando que la suma de sus componentes sea 1.
2. **Conjugación:** La distribución de Dirichlet es conjugada previa de la distribución multinomial, permitiendo una actualización eficiente de las creencias a partir de nuevos datos.
3. **Esperanza y Varianza:**

- La esperanza de cada componente x_i de un vector \mathbf{x} distribuido según Dirichlet es:

$$\mathbb{E}[x_i] = \frac{\alpha_i}{\sum_{j=1}^K \alpha_j}$$

- La varianza de cada componente x_i es:

$$\text{Var}(x_i) = \frac{\alpha_i(\alpha_0 - \alpha_i)}{\alpha_0^2(\alpha_0 + 1)}$$

donde $\alpha_0 = \sum_{i=1}^K \alpha_i$.

En esta investigación, utilizaremos Latent Dirichlet Allocation (LDA) para el análisis de tópicos en conjuntos de textos. LDA es un modelo generativo que permite descubrir la estructura subyacente de los datos textuales mediante la identificación de tópicos latentes en un corpus. Este enfoque nos ayudará a comprender mejor las distribuciones de palabras y tópicos en los documentos analizados.

3 LDA (Latent Dirichlet Allocation)

3.1 Presentación del modelo LDA

Iniciamos exponiendo las nociones fundamentales que sustentan LDA, considerado el modelo de tópicos más elemental. LDA se fundamenta en la premisa de que los documentos incorporan múltiples tópicos. Tomemos, por ejemplo, el artículo ilustrado en la Figura 1, extracto de la propuesta suscrita por Eugenio Arcos, diputado de la legislatura LXII del partido movimiento ciudadano. “Iniciativa con proyecto a decreto que insta adiconar el articulo 48, de la ley de ingresos”. El diputado propone modificar la Ley de Ingresos de Ahualulco para establecer sanciones por el incumplimiento del uso obligatorio de cubrebocas y medidas sanitarias durante la pandemia de COVID-19, en alineación con las recomendaciones de la OMS y la Secretaría de Salud federal, para mitigar los riesgos de salud pública en San Luis Potosí.

Manualmente diversas palabras empleadas en el artículo se han subrayado. Términos relacionados con el presupuesto del estado, como ingresos y fiscal, se marcan en celeste; términos vinculados a la administración pública, como leyes y municipios, en rosa; y aquellos referentes a la pandemia, como sanitarias y COVID-19, en amarillo. Si procediéramos a resaltar cada término relevante en el texto, observaríamos cómo este artículo entrelaza tópicos sobre covid-19, presupuesto del estado y administración pública en distintas proporciones. (Se han omitido palabras con escaso valor temático como “y”, “de” o “en”, etc). Reconocer la intersección de estos tópicos en el artículo facilitaría su clasificación dentro de un corpus de propuestas legislativas.

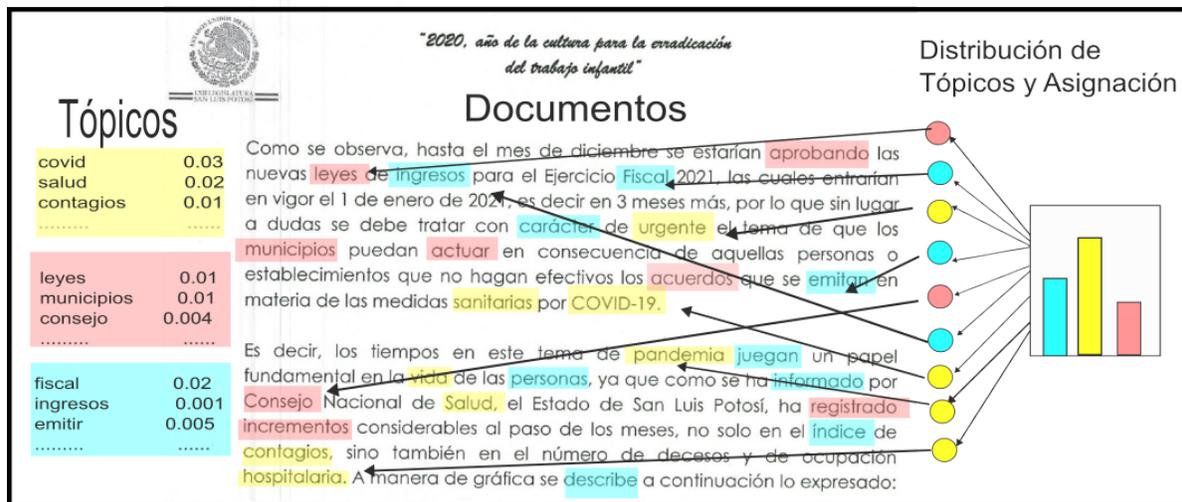


Figura 1: Proceso generador ejemplificado una propuesta legislativa del Congreso de S.L.P.

LDA se basa en la suposición de bolsa de palabras D. M. Blei, Ng, y Jordan (2003), lo que significa que las palabras en un documento son intercambiables y su orden por lo tanto no es

importante. Esto lleva a la representación de una colección de documentos como una llamada matriz documento-término (DTM Document Term Matrix por sus siglas en inglés), en la que se capturan las frecuencias de las palabras en los documentos. La Tabla 2 muestra una matriz documento-término de ejemplo para un pequeño corpus.

Tabla 2: Matriz documento-término ejemplo con $D = 3$ documentos y vocabulario de tamaño $V = 5$ términos.

	Término 1	Término 2	Término 3	Término 4	Término 5=V
Documento 1	2	1	0	0	1
Documento 2	1	0	2	3	1
Documento 3=D	0	2	4	0	0

D. M. Blei (2012) formalmente, define un tópico como una distribución sobre un vocabulario fijo. Por ejemplo, el tópico de COVID-19 tiene palabras sobre la pandemia con alta probabilidad, y el tópico sobre el presupuesto del estado tiene palabras relacionadas con el fisco, impuestos, dinero con alta probabilidad (el modelo asume que estos tópicos se especifican antes de que se haya generado cualquier dato).

En la Figura 1, de ejemplo, la distribución sobre tópicos asignaría probabilidades a COVID-19, presupuesto del estado y administración pública. Cada palabra proviene de uno de esos tres tópicos. Es importante notar que el siguiente documento en el corpus podría ser sobre pandemia y economía, u otros tópicos. Su distribución sobre tópicos asignaría probabilidades a esos dos tópicos. Esta es la característica distintiva de la asignación de Dirichlet latente: todos los documentos en el corpus comparten el mismo conjunto de tópicos, pero cada documento exhibe esos tópicos en diferentes proporciones D. M. Blei (2012).

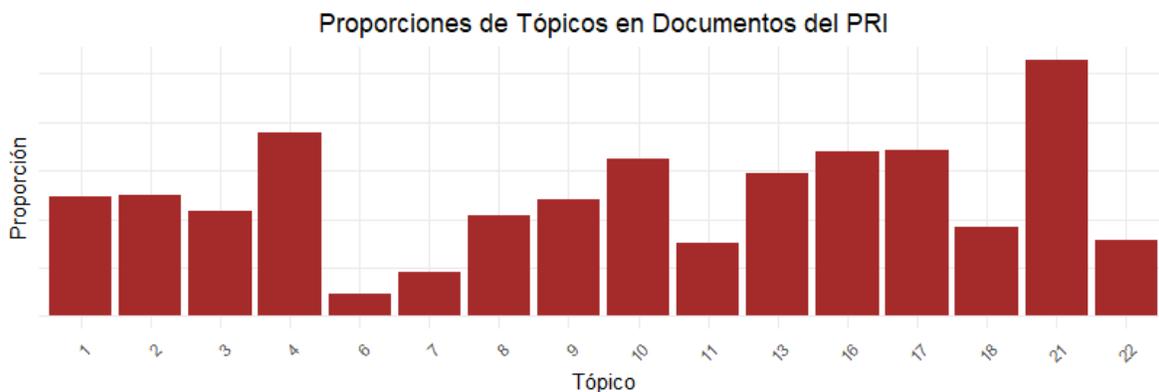


Figura 2: Ejemplo de Distribución de Tópicos en los documentos propuestos por diputados del PRI

La Figura 2, muestra un ejemplo de inferencia utilizando el corpus de documentos de las tres legislaturas. Aquí seleccionamos más de 3600 propuestas del Congreso de San Luis Potosí y aplicamos un algoritmo de modelado de tópicos para deducir la estructura latente de estos (el algoritmo suponía la existencia de 25 tópicos¹). Luego, calculamos la distribución de tópicos inferida para las propuestas legislativas del partido político PRI, es decir, la distribución a través de los tópicos que mejor describen su colección particular de palabras. Observamos que esta distribución de tópicos puede utilizar cualquiera de los tópicos disponibles. Además, podemos revisar los términos más probables de cada uno de los tópicos más destacados en la Figura 3. Al examinarlos, notamos que estos términos se reconocen como relacionados con la Sanitario-COVID, administrativos y el presupuesto, los cuales se combinan en el artículo de ejemplo.

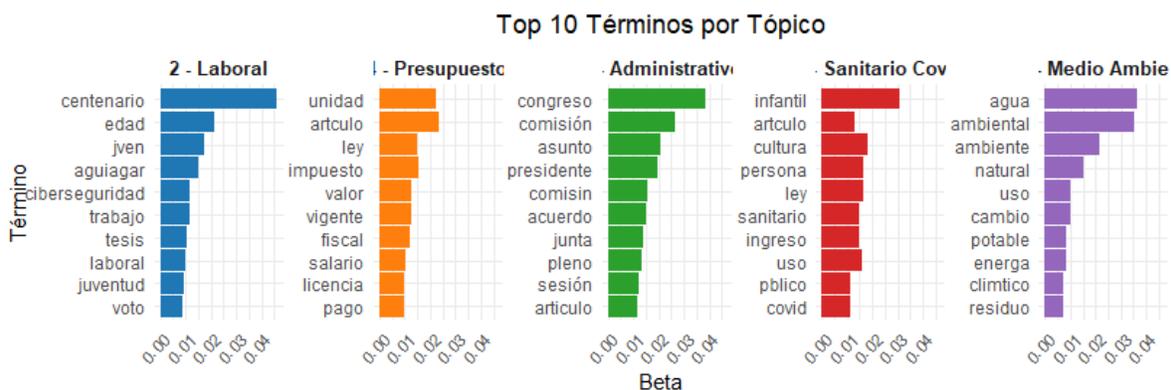


Figura 3: Términos por tópico de las propuestas de diputados del PRI

El objetivo del modelado de tópicos es descubrir automáticamente los tópicos a partir de una colección de documentos (en nuestro caso las propuestas del congreso). Los documentos son observados, mientras que la estructura de tópicos (los tópicos, las distribuciones de tópicos por documento y las asignaciones de tópicos por documento y palabra) es latente. El problema central del modelado de tópicos es inferir esta estructura latente a partir de los documentos observados, lo cual se puede considerar como “invertir” el proceso generativo (ver Figura 4): ¿cuál es la estructura latente que probablemente generó la colección observada?

La utilidad de los modelos de tópicos radica en que la estructura latente inferida se asemeja a la estructura temática de la colección. Esta estructura latente, que es interpretable, anota cada documento en la colección, una tarea laboriosa de realizar manualmente. Estas anotaciones pueden utilizarse para ayudar en tareas como la recuperación de información, clasificación y exploración del corpus Griffiths y Steyvers (2004). De esta manera, el modelado de tópicos proporciona una solución algorítmica para gestionar, organizar y anotar grandes archivos de textos.

¹Consulte la Sección 7.2 para conocer el método de cálculo de K.

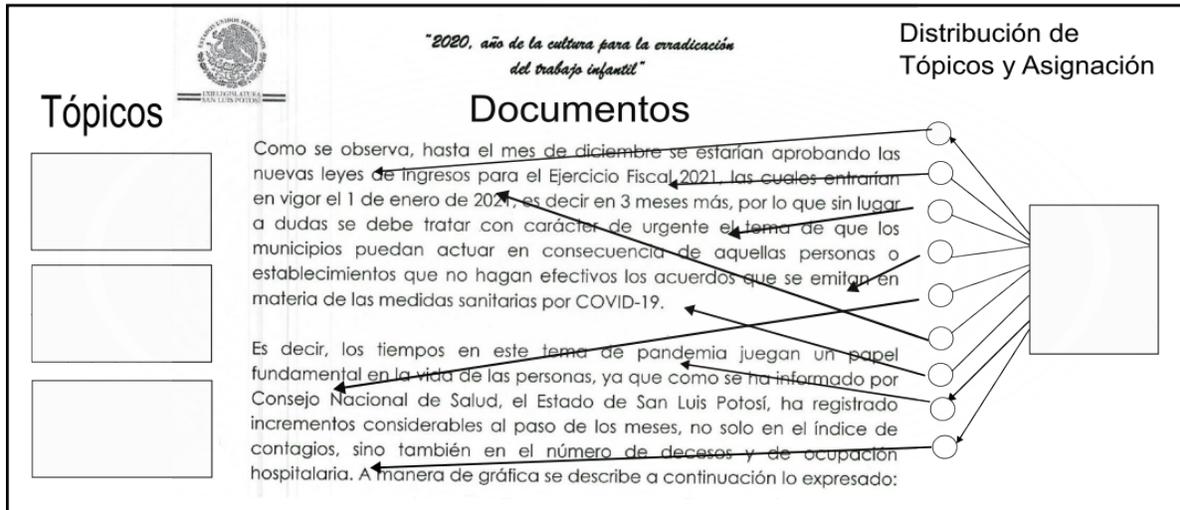


Figura 4: Proceso inverso de la Figura 1

3.2 Proceso Generativo

LDA y otros modelos de tópicos pertenecen al campo más amplio del modelado probabilístico. En el modelado probabilístico generativo, consideramos nuestros datos como si surgieran de un proceso generativo que incluye variables latentes. Este proceso define una distribución de probabilidad conjunta para tanto las variables aleatorias observadas como las latentes. Utilizamos esta distribución conjunta para calcular la distribución condicional de las variables latentes dadas las observadas, también conocida como distribución posterior D . M. Blei (2012). LDA se enmarca precisamente en este contexto. Las variables observadas son las palabras de los documentos; las variables latentes son la estructura tópica; y el proceso generativo es como se describe siguiendo a D. M. Blei, Ng, y Jordan (2003). Blei propone el siguiente proceso generativo para cada documento w dentro de un corpus D :

1. Determinar N a partir de una distribución Poisson(ξ).
2. Seleccionar θ siguiendo una distribución Dir(α).
3. Para cada una de las N palabras w_n :
 - a. Escoger un tópico z_n de acuerdo con una distribución Multinomial(θ).
 - b. Seleccionar una palabra w_n basándose en $p(w_n|z_n, \beta)$, que es la probabilidad multinomial condicional al tópico z_n .

Siguiendo a Ponweiser (2012) el proceso generativo para LDA corresponde a la siguiente distribución conjunta de las variables ocultas y observadas.

$$p(w, z, \theta, \phi | \alpha, \beta) = p(\theta | \alpha)p(z | \theta)p(\phi | \beta)p(w | z, \phi) \quad (1)$$

Integrar sobre la distribución conjunta nos permitirá obtener las probabilidades buscadas que están contenidas en el modelo. Se explicará los cuatro factores en la parte derecha de la Ecuación 1.

Primero, la distribución de tópicos por documento, que se extrae de la distribución de Dirichlet, dado el parámetro de Dirichlet α , que es un vector K con componentes $\alpha_k > 0$:

$$p(\theta | \alpha) = \frac{\Gamma\left(\sum_{k=1}^K \alpha_k\right)}{\prod_{k=1}^K \Gamma(\alpha_k)} \theta_1^{\alpha_1-1} \dots \theta_K^{\alpha_K-1} \quad (2)$$

Tabla 3: Ejemplo de asignación de tópicos por documento.

		Tópico 1	Tópico 2	Tópico 3	Tópico 4
Documento 1	$\theta_{d=1}$	0.5	0.01	0.3	0.1
Documento 2	$\theta_{d=2}$	0.01	0.09	0.1	0.04
Documento 3	$\theta_{d=3=D}$	0.02	0.48	0.25	0.25

El siguiente componente de la distribución conjunta es la distribución de las asignaciones de tópico a palabras en el corpus, z , que depende de la distribución mencionada anteriormente θ . Por lo tanto, a cada palabra w_i en un documento de N palabras se le asigna un valor de $1 \dots K$. La siguiente tabla muestra un ejemplo de esto:

Tabla 4: Ejemplo de asignación de tópicos k a términos w_i (palabras) en el corpus.

		w_1	w_2	w_3	w_4	w_5	w_6
Documento 1	$z_{d=1}$	k = 2	k = 1	k = 1	k = 4	k = 3	k = 3
Documento 2	$z_{d=2}$	k = 2	k = 3	k = 2	k = 2	k = 2	k = 2
Documento 3	$z_{d=3=D}$	k = 4	k = 2	k = 2	k = 4	k = 3	k = 1

En la distribución conjunta expresamos la probabilidad de z para todos los documentos y tópicos en términos del conteo de palabras $n_{d,k}$, que es el número de veces que el tópico k ha sido asignado a cualquier palabra en el documento d :

$$p(z | \theta) = \prod_{d=1}^D \prod_{k=1}^K \theta_{d,k}^{n_{d,k}} \quad (3)$$

Las distribuciones de términos por tópico de todo el corpus ϕ_k son (de nuevo) extraídas de una distribución de Dirichlet con el parámetro β . Este es un ejemplo de algunas posibles distribuciones de tópicos para cuatro tópicos donde el término v sea extraído cuando se eligió el tópico k .

Tabla 5: Ejemplo de asignación de términos (palabras) por tópico.

		Término 1	Término 2	Término 3	Término 4	Término 5=V
Tópico 1	$\phi_{k=1}$	0.1	0.1	0	0.7	0.1
Tópico 2	$\phi_{k=2}$	0.2	0.1	0.2	0.2	0.3
Tópico 3	$\phi_{k=3}$	0.01	0.2	0.39	0.3	0.1
Tópico 4	$\phi_{k=4=k}$	0.0	0.0	0.5	0.3	0.2

Ahora expresamos la probabilidad de ϕ para todos los tópicos y todas las palabras del vocabulario como:

$$p(\phi | \beta) = \prod_{k=1}^K \frac{\Gamma(\beta_{k,\cdot})}{\prod_{v=1}^V \Gamma(\beta_{k,v})} \prod_{v=1}^V \phi_{k,v}^{\beta_{k,v}-1} \quad (4)$$

Finalmente, la probabilidad de un corpus w dado sus predecesores z y ϕ :

$$p(w | z, \phi) = \prod_{k=1}^K \prod_{v=1}^V \phi_{k,v}^{n_{r,k,v}} \quad (5)$$

donde $n_{r,k,v}$ es el conteo de cuántas veces el tópico k fue asignado al término del vocabulario v en todo el corpus. Luego reemplazando Ecuación 2, Ecuación 3, Ecuación 4 y Ecuación 5 en Ecuación 1 tenemos la distribución conjunta completa y reagrupada:

$$\begin{aligned} p(w, z, \theta, \phi | \alpha, \eta) &= p(\theta | \alpha) p(z | \theta) p(\phi | \beta) p(w | z, \phi) = \\ &= \left(\prod_{d=1}^D \frac{\Gamma(\alpha_{\cdot})}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K \theta_{d,k}^{\alpha_k-1} \right) \left(\prod_{d=1}^D \prod_{k=1}^K \theta_{d,k}^{n_{d,k,\cdot}} \right) \times \\ &\quad \left(\prod_{k=1}^K \frac{\Gamma(\beta_{k,\cdot})}{\prod_{v=1}^V \Gamma(\beta_{k,v})} \prod_{v=1}^V \phi_{k,v}^{\beta_{k,v}-1} \right) \left(\prod_{k=1}^K \prod_{v=1}^V \phi_{k,v}^{n_{r,k,v}} \right) \\ &= \left(\prod_{d=1}^D \frac{\Gamma(\alpha_{\cdot})}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K \theta_{d,k}^{\alpha_k+n_{d,k,\cdot}-1} \right) \times \\ &\quad \left(\prod_{k=1}^K \frac{\Gamma(\beta_{k,\cdot})}{\prod_{v=1}^V \Gamma(\beta_{k,v})} \prod_{v=1}^V \phi_{k,v}^{\beta_{k,v}+n_{r,k,v}-1} \right). \end{aligned} \quad (6)$$

Ahora siguiendo a Ponweiser (2012) marginalizamos las variables latentes para poder escribir la probabilidad del modelo cuando se da un corpus w y los hiperparámetros (α y β). Esta probabilidad es necesaria para una “estimación de máxima verosimilitud de los parámetros del modelo y para inferir la distribución de las variables latentes”:

$$p(w | \alpha, \beta) = \int_{\phi} \int_{\theta} \sum_z \left(\prod_{d=1}^D \frac{\Gamma(\alpha_{\cdot})}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K \theta_{d,k}^{\alpha_k + n_{d,k} - 1} \right) \times \left(\prod_{k=1}^K \frac{\Gamma(\beta_{k\cdot})}{\prod_{v=1}^V \Gamma(\beta_{k,v})} \prod_{v=1}^V \phi_{k,v}^{\beta_{k,v} + n_{\cdot,k,v} - 1} \right) d\theta d\phi. \quad (7)$$

La suma sobre todas las posibles combinaciones de asignaciones de tópico z ($n_{d,k}, n_{\cdot,k,v}$) hace que esta probabilidad sea computacionalmente intratable D. M. Blei, Ng, y Jordan (2003). De forma más técnica, la suma abarca todas las posibles maneras de asignar cada palabra observada de la colección a uno de los tópicos. Las colecciones de documentos generalmente contienen palabras observadas, al menos, en el orden de los millones. En su lugar, tenemos que recurrir a técnicas computacionales para encontrar buenas aproximaciones de la probabilidad marginal.

4 Estimación de LDA mediante Muestreo de Gibbs

El problema computacional de inferir la estructura tópica oculta a partir de los documentos consiste en calcular la distribución posterior, es decir, la distribución condicional de las variables latentes dadas los documentos (David M. Blei y Jordan 2006). Estimar los parámetros para LDA de toda la colección de datos en Ecuación 7 es intratable. La solución a esto es utilizar métodos de estimación aproximada como los Métodos Variacionales (D. M. Blei, Ng, y Jordan 2003), la Propagación de Expectativas (Minka y Lafferty 2002) y el Muestreo de Gibbs (Griffiths y Steyvers 2004). En este trabajo utilizaremos el Muestreo de Gibbs, que es un caso especial de los métodos de Monte Carlo con cadenas de Markov (MCMC) (Geman y Geman 1984). Usamos este método debido a que según Heinrich (2009) produce algoritmos relativamente simples para la inferencia aproximada en modelos de alta dimensionalidad como LDA .

El primer uso del Muestreo de Gibbs² para estimar LDA se reporta en Griffiths y Steyvers (2004). Las principales variables de interés en el modelo son las distribuciones de tópico-palabra ϕ y las distribuciones de tópicos θ para cada documento. Siguiendo a (Steyvers y Griffiths 2004; Griffiths, Thomas y Steyvers, Mark 2007) en la cual representan la colección de documentos por un conjunto de índices de palabras w_i e índices de documentos d_i , para cada token de palabra i . El procedimiento de muestreo de Gibbs considera cada token de palabra en la colección de texto por turno, y estima la probabilidad de asignar el token de palabra actual a cada tema, condicionado a las asignaciones de tema a todos los demás tokens de palabra. A partir de esta distribución condicional, se elige un tema y se almacena como la nueva asignación de tema para este token de palabra. Escribimos esta distribución condicional como $P(z_i = j | \mathbf{z}_{-i}, w_i, d_i, \cdot)$, donde $z_i = j$ representa la asignación de tema del token i al tema j , \mathbf{z}_{-i} se refiere a las asignaciones de tema de todos los demás tokens de palabra, y “ \cdot ” se refiere a toda otra información conocida u observada, como todos los demás índices de palabras y documentos w_{-i} y d_{-i} , y los hiperparámetros α y β . Griffiths y Steyvers (2004) mostraron que esta distribución puede ser calculada como:

$$P(z_i = j | \mathbf{z}_{-i}, w_i, d_i, \cdot) \propto \frac{C_{w_i j}^{WT} + \beta}{\sum_{w=1}^W C_{w j}^{WT} + W\beta} \frac{C_{d_i j}^{DT} + \alpha}{\sum_{t=1}^T C_{d_i t}^{DT} + T\alpha} \quad (8)$$

Donde \mathbf{C}^{WT} y \mathbf{C}^{DT} son matrices de conteo con dimensiones $W \times T$ y $D \times T$, respectivamente;

$C_{w j}^{WT}$ contiene el número de veces que la palabra w se asigna al tema j , sin incluir la instancia actual i , y $C_{d j}^{DT}$ contiene el número de veces que el tema j se asigna a algún token de palabra en el documento d , sin incluir la instancia actual i .

La probabilidad real de asignar un token de palabra al tema j se calcula dividiendo $P(z_i = j | \mathbf{z}_{-i}, w_i, d_i, \cdot)$ para el tema t entre la suma de todos los tópicos T .

²Una descripción más completa de este método se encuentra en el informe técnico Heinrich (2009)

Griffiths y Steyvers (2004) dividen la Ecuación 8 en dos partes (Derecha - Izquierda). La parte izquierda es la probabilidad de la palabra w bajo el tema j , mientras que la parte derecha es la probabilidad que el tema j tiene bajo la distribución actual de tópicos para el documento d . Una vez que muchos tokens de una palabra se han asignado al tema j (a lo largo de varios documentos), aumentará la probabilidad de asignar cualquier token de esa palabra en particular al tema j . Al mismo tiempo, si el tema j se ha utilizado múltiples veces en un documento, aumentará la probabilidad de que cualquier palabra de ese documento se asigne al tema j . Por lo tanto, las palabras se asignan a tópicos según cuán probable sea que la palabra pertenezca a un tema y cuán dominante sea un tema en un documento.

4.0.1 Estimación de Parámetros (ϕ y θ).

El algoritmo de muestreo proporciona estimaciones directas de z para cada palabra. Sin embargo, muchas aplicaciones del modelo requieren estimaciones ϕ y θ de las distribuciones término-tópico y tópico-documento respectivamente. Estas se pueden obtener a partir de las matrices de conteo de la siguiente manera³:

$$\phi_{ij} = \frac{C_{ij}^{WT} + \beta}{\sum_{k=1}^W C_{kj}^{WT} + W\beta} \quad \theta_{dj} = \frac{C_{dj}^{DT} + \alpha}{\sum_{k=1}^T C_{dk}^{DT} + T\alpha} \quad (9)$$

Estos valores corresponden a las distribuciones predictivas de muestrear un nuevo token de la palabra i del tema j , y muestrear un nuevo token (aún no observado) en el documento d del tema j , y también son las medias posteriores de estas cantidades condicionadas a una muestra particular \mathbf{z} . (Griffiths, Thomas y Steyvers, Mark 2007; Griffiths y Steyvers 2004; Mark Steyvers, Tom Griffiths 2007)

³Para ver un ejemplo práctico de la programación para la obtención de las matrices ϕ y θ , consulte la Sección 7.5.

5 LDA aplicado a las Propuestas Legislativas del Congreso de San Luís Potosí

El proceso realizado con las propuestas legislativas del Congreso de San Luís Potosí incluye varias etapas importantes y la utilización de diversos paquetes de R resumidos en la Figura 5.

5.1 Descripción del Proceso

1. Digitalización y OCR⁴:

- **Web Scraping:** Se descargaron las propuestas legislativas desde el sitio web del congreso utilizando técnicas de web scraping (Congreso, SLP 2023).
- **OCR (Reconocimiento Óptico de Caracteres):** Dado que las propuestas legislativas se publican como documentos PDF escaneados, se utilizó OCR para convertir las imágenes de texto en texto editable (Ooms 2022).

2. Filtrado y Limpieza de Texto⁵:

Definición 5. Palabras sin sentido: Son aquellas que, debido a la calidad de los documentos y las limitaciones del OCR, no son extraídas correctamente. Estas palabras surgen de ruidos o mala calidad de la imagen del documento original, resultando en términos que no existen en el idioma español. Por ejemplo, “polilico” debería ser “político”.

Esto nos ocurrió en las propuestas legislativas del Congreso de San Luis Potosí debido a que los documentos los suben en mal estado. Observamos que estos documentos escaneados son de baja calidad, sumado a factores como el envejecimiento del papel, manchas, dobleces, sombras y tipos de letra poco claros, lo que nos causo muchos errores. Para mitigar estos problemas, usamos técnicas de preprocesamiento de imágenes (ajuste de contraste, eliminación de ruido), postprocesamiento de texto (corrección ortográfica) y revisión manual. Estas estrategias mejoraron la precisión del OCR, crucial para la digitalización precisa y útil de documentos. Es así que se creó un diccionario específico para filtrar estas palabras sin sentido. Este diccionario se utilizó para limpiar el texto de las propuestas.

3. Tokenización y Stopwords:

Definición 6. Stopwords: Son palabras comunes en un idioma que suelen tener poco valor semántico por sí mismas, como “el”, “la”, “y”, “de”, entre otras. En el análisis de textos, estas palabras se eliminan porque no aportan información relevante sobre el contenido o el tema del texto. La eliminación de stopwords ayuda a reducir el ruido en los datos y mejora la precisión de los modelos de análisis de texto (D. M. Blei 2012).

⁴Para obtener detalles sobre el proceso de digitalización de las legislaturas, consulte la Sección 7.3.

⁵Para mayor información del proceso de LDA ir la Sección 7.3.

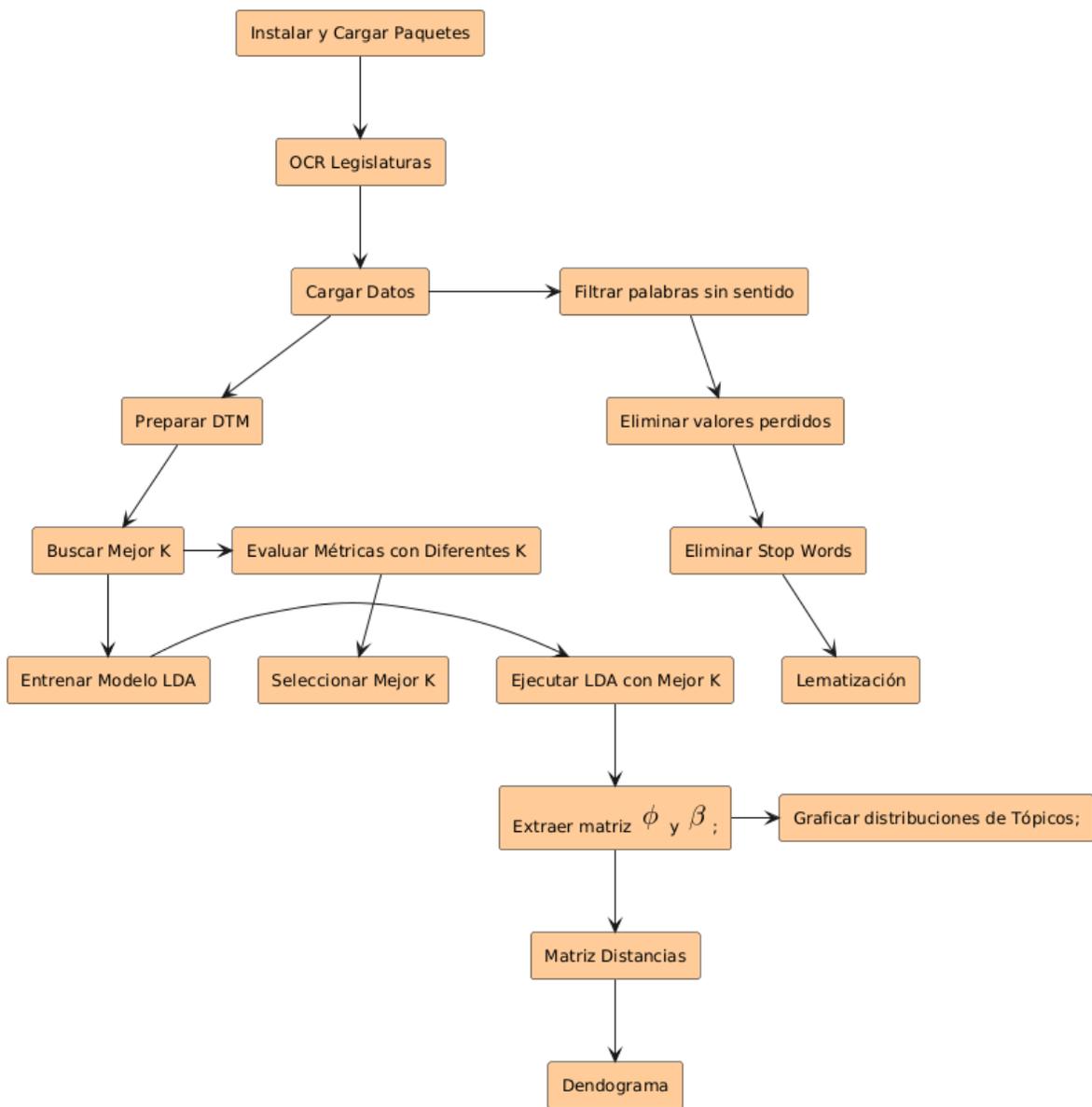


Figura 5: Diagrama de Trabajo

Definición 7. Tokenización: Es el proceso de dividir un texto en unidades más pequeñas llamadas “tokens”. Estos tokens pueden ser palabras, frases u otros elementos significativos que faciliten el análisis. Por ejemplo, el texto “La casa es grande” se tokeniza en [“La”, “casa”, “es”, “grande”]. La tokenización es un paso esencial en el preprocesamiento de texto porque permite que el análisis sea manejable y estructurado (D. M. Blei 2012). Además se integraron diccionarios de stopwords en español de varias fuentes para eliminar palabras comunes y no informativas del texto (Benoit et al. 2018, 2023).

4. Enriquecimiento de la Base de Datos:

Definición 8. Metadata Manual: Se agregó manualmente a la base de datos el género y el partido político del proponente de cada propuesta legislativa. Esta información nos será útil para conocer la distribución de tópicos por partido político y para el análisis de clúster de partidos políticos.

5. Análisis de Modelos de Temas (LDA):

Paquetes de R Utilizados:

- `topicmodels` para el modelado LDA.
- `tidytext`, `quanteda`, `udpipe`, `stopwords` para la preparación y limpieza del texto.
- `ggplot2`, `ldatuning` para la visualización y optimización del número de temas (K) en los modelos LDA (Grün y Hornik 2023a; J. Chang 2015; Nikita 2020; Wickham 2016).

6. **Determinación del Mejor K:** Se utilizó el paquete `ldatuning` para encontrar el número óptimo de temas, evaluando varios valores de K y seleccionando el mejor basado en métricas específicas⁶.

7. **Modelo LDA:** Se ajustó el modelo LDA con el número de temas óptimo K, utilizando el método de muestreo de Gibbs para identificar los temas presentes en las propuestas legislativas.

⁶En nuestro caso utilizamos el valor de $k=25$, para mayor información del proceso de obtención de k consulte la Sección 7.2.

6 Análisis de Resultados

6.1 Base de Datos

Exploramos 3,645 documentos pdf (propuestas legislativas) presentadas entre enero de 2016 y el diciembre de 2022 tomadas del Congreso San Luis Potosí (2023). La base de datos utilizada en esta tesis contiene información detallada con un total de 39,919,272 caracteres y 6,692,905 palabras.

Tabla 6: Resumen de la cantidad de propuestas legislativas por proponente y sexo

Proponente	Sexo	Cantidad de Propuestas
CIUDADANO	F	21
CIUDADANO	M	129
COALICIÓN	F	1
GOBERNADOR	M	104
INDEPENDIENTES	M	3
MORENA	F	301
MORENA	M	179
PAN	F	218
PAN	M	427
PCP	F	22
PCP	M	151
PES	M	19
PMC	F	73
PMC	M	96
PNA	F	122
PNA	M	11
PR	F	1
PRD	F	126
PRD	M	114
PRI	F	351
PRI	M	384
PT	F	21
PT	M	36
EMPRESA PÚBLICA	F	10
EMPRESA PÚBLICA	M	37
PVEM	F	15
PVEM	M	259
RSP	F	40

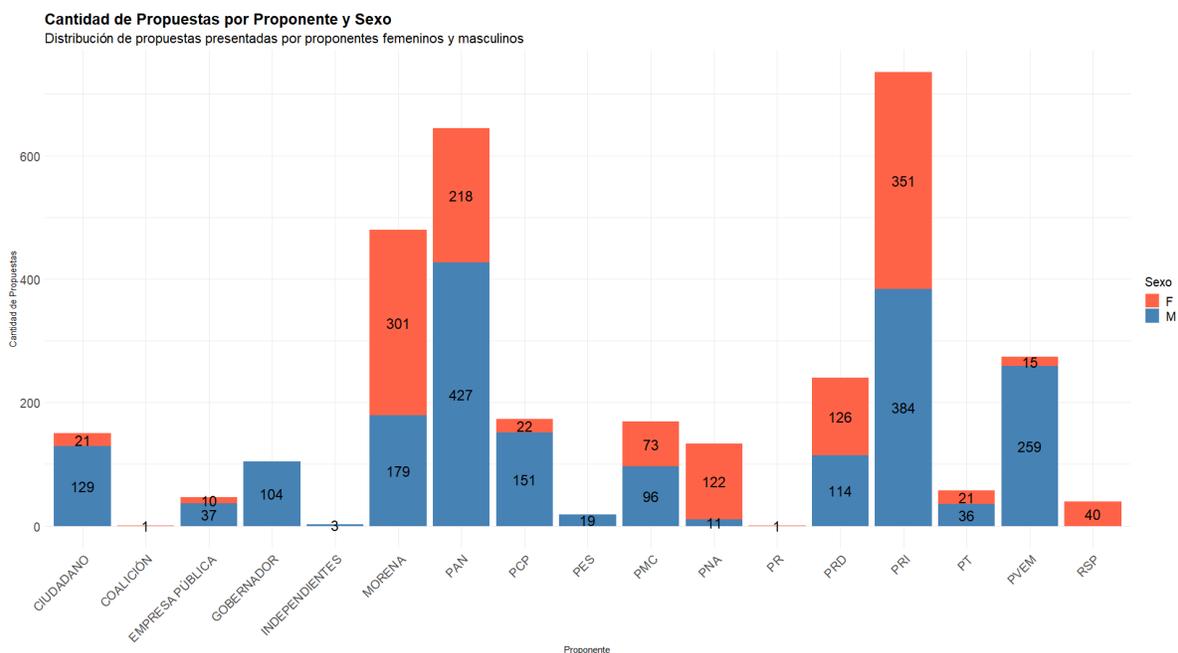


Figura 6: Cantidad de Propuestas por Partido y Sexo

Distribución de Géneros en los Partidos

Este análisis se centran en la distribución de géneros en los diferentes partidos, no en los hallazgos clave de la tesis en general (Figura 6 y Tabla 6):

1. PRI (Partido Revolucionario Institucional):

- Lidera en participación femenina con 351 propuestas, destacándose como el partido con mayor actividad legislativa de mujeres.
- Los hombres en el PRI presentaron 384 propuestas, mostrando un balance relativamente equilibrado entre géneros.

2. MORENA (Movimiento Regeneración Nacional):

- Ocupa el segundo lugar con 301 propuestas presentadas por mujeres.
- La diferencia de género es menor en MORENA, con 179 propuestas presentadas por hombres, sugiriendo una mayor equidad de género.

3. PAN (Partido Acción Nacional):

- Las mujeres presentaron 218 propuestas, situando al PAN en el tercer lugar.
- Existe una mayor disparidad de género en el PAN, ya que los hombres presentaron 427 propuestas.

4. Otros Partidos:

- Partidos como el PRD (126 propuestas) y el PNA (122 propuestas) también muestran una participación femenina considerable.
- COALICIÓN y PR tienen una participación femenina mínima, con solo 1 propuesta cada uno.

En resumen, el PRI se destaca por su alta actividad legislativa femenina, seguido por MORENA y PAN. Aunque algunos partidos muestran un balance de género más equitativo, otros presentan una notable disparidad, indicando la necesidad de mayores esfuerzos para promover la participación de las mujeres en la política legislativa.

6.2 Análisis de Tópicos Latentes en Datos del Congreso de S.L.P.

En este trabajo, utilizamos el software R, específicamente el paquete `topicmodels` para el componente de modelado de tópicos. LDA asume que los temas son los mismos para todos los documentos, y solo las proporciones de los tópicos varían. Por lo tanto, `topicmodels` requiere una entrada que especifique el número de tópicos a descubrir. Elegir este número es crítico para el éxito de un modelo de tópicos, ya que muy pocos temas pueden fusionar distintos tópicos, mientras que demasiados tópicos pueden introducir muchos “tópicos” que consisten en vocabularios que parecen no tener nada en común, o incluso comenzar a dividir tópicos que eran identificables con valores de entrada más pequeños (Curran et al. 2018; Grün y Hornik 2023b, 2023a).

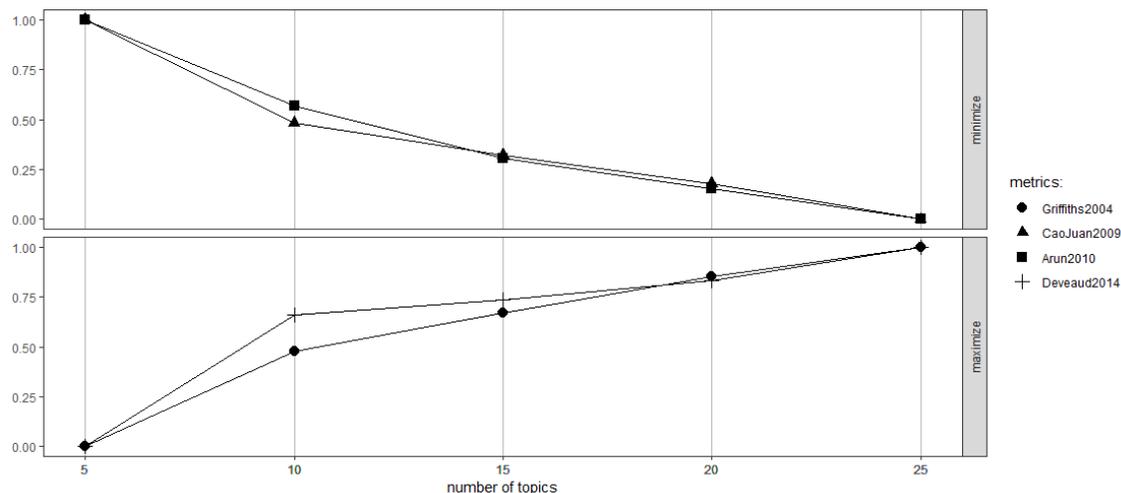


Figura 7: Número óptimo de tópicos con diferentes Métricas de Evaluación

Para nuestro análisis, determinamos que el valor óptimo de k es 25 utilizando cuatro métricas (ver Figura 7): Griffiths y Steyvers (2004), que maximiza la probabilidad de los datos mediante muestreo de Gibbs colapsado; Cao et al. (2009), que minimiza la distancia media entre distribuciones de tópicos; Arun et al. (2010), que usa la divergencia KL entre distribuciones de tópicos

y documentos; y Deveaud, SanJuan, y Bellot (2014), que maximiza la coherencia de tópicos midiendo la similitud de las palabras más probables. Además, es fundamental que los temas sean fácilmente identificables y distintos entre sí. Nosotros encontramos que 25 tópicos satisfacen estos requisitos, equilibrando robustez probabilística, interpretación clara y coherencia temática. Los tópicos identificados y sus palabras clave correspondientes se pueden encontrar en la Tabla 7 y Figura 8. Sin embargo, al revisar manualmente una muestra aleatoria de los tópicos para asignarles etiquetas, identificamos algunos temas que, por ser suficientemente similares, decidimos consolidarlos. Así, agrupamos los tópicos de la siguiente manera:

- **Urbanismo:** unimos los tópicos 1 y 24.
- **Presupuesto:** unimos los tópicos 4 y 15.
- **Administrativos:** unimos los tópicos 5, 12 y 20.
- **Educación y Cultura:** unimos los tópicos 9 y 19.
- **Protección y Asistencia a Víctimas:** unimos los tópicos 17 y 25.
- **Basura:** El tópico 14 se eliminó debido a que contenía palabras irrelevantes, lo cual es común dado que los datos se obtuvieron mediante un proceso de OCR con información ruidosa.

Esta consolidación nos permitió reducir la redundancia y mejorar la claridad de los tópicos para un análisis más preciso y coherente.

Tabla 7: Palabras clave probabilísticas para etiquetar los tópicos en propuestas del Congreso de S.L.P.

Tópico y Etiqueta	Términos
1. Urbanismo	indgena, pueblo, comunidad, propiedad, indgeno, calle, ayuntamiento,...
2. Laboral	centenario, edad, joven, aguiagar, ciberseguridad, trabajo, laboral,...
3. Electoral	electoral, político, ciudadano, candidato, artículo, consejo, ley,...
4. Presupuesto	artículo, unidad, impuesto, ley, valor, vigente, fiscal, salario,...
5. Administrativos	congreso, comisión, asunto, presidente, acuerdo, junta, pleno, sesión,...
6. Sanitario COVID	infantil, cultura, ley, persona, uso, sanitario, ingreso, público, covid,....
7. Agricultura y Ganadería	rural, producto, productor, sustentable, vegetal, actividad, agrícola, pesca,....
8. Familia	familiar, adolescente, hijo, familia, civil, registro, padre,...
9. Educación y Cultura	cultural, cultura, patrimonio, paz, turístico, turismo, arte,...
10. Leyes de tránsito y Transporte Público	urbano, vehículo, seguridad, público, territorial, uso, vivienda, ordenamiento,...
11. Comunicación	información, archivo, acceso, transparencia, electrónico, sujeto, obligado,....
12. Administrativos	ley, artículo, congreso, política, establecer, adicionar, quedar, público,...
13. Minorías Sociales	persona, humano, violencia, género, social, igualdad, vida, acceso,....
14. Basura	ea, rr, ia, per, rn, oa, cr, eer, fa, lr,...
15. Presupuesto	presupuesto, público, gasto, ejercicio, fiscal, económico, ingreso, recurso,...
16. Auditoría	público, ley, superior, congreso, administrativo, responsabilidad, auditora,...
17. Protección y asistencia a Víctimas	seguridad, víctima, general, público, ley, justicia, humano, persona, penal,...
18. Medio Ambiente	agua, ambiental, ambiente, natural, uso, cambio, potable, energía, climático,...
19. Educación y Cultura	educativo, escolar, cultura, superior, escuela, profesional, estudio, nivel,...

Tópico y Etiqueta	Términos
20. <i>Administrativos</i>	artículo, procedimiento, tribunal, administrativo, judicial, autoridad, acto,...
21. <i>Justicia Social y desarrollo inclusivo</i>	ley, artículo, estatal, pblico, programa, social, general, autoridad,...
22. <i>Salud Comunitaria</i>	salud, sanitario, contingencia, persona, colaborar, solidaridad, civil, covid,...
23. <i>Leyes de transito y Transporte Publico</i>	transporte, pblico, femenino, nieto, sufragio, universitario, promotor, vehculo,...
24. <i>Urbanismo</i>	pblico, bien, contrato, financiamiento, pago, ley, crdito, obra, obligación,...
25. <i>Protección y asistencia a Víctimas</i>	violencia, penal, pena, artículo, cdigo, persona, sexual, conducta, vctima,...

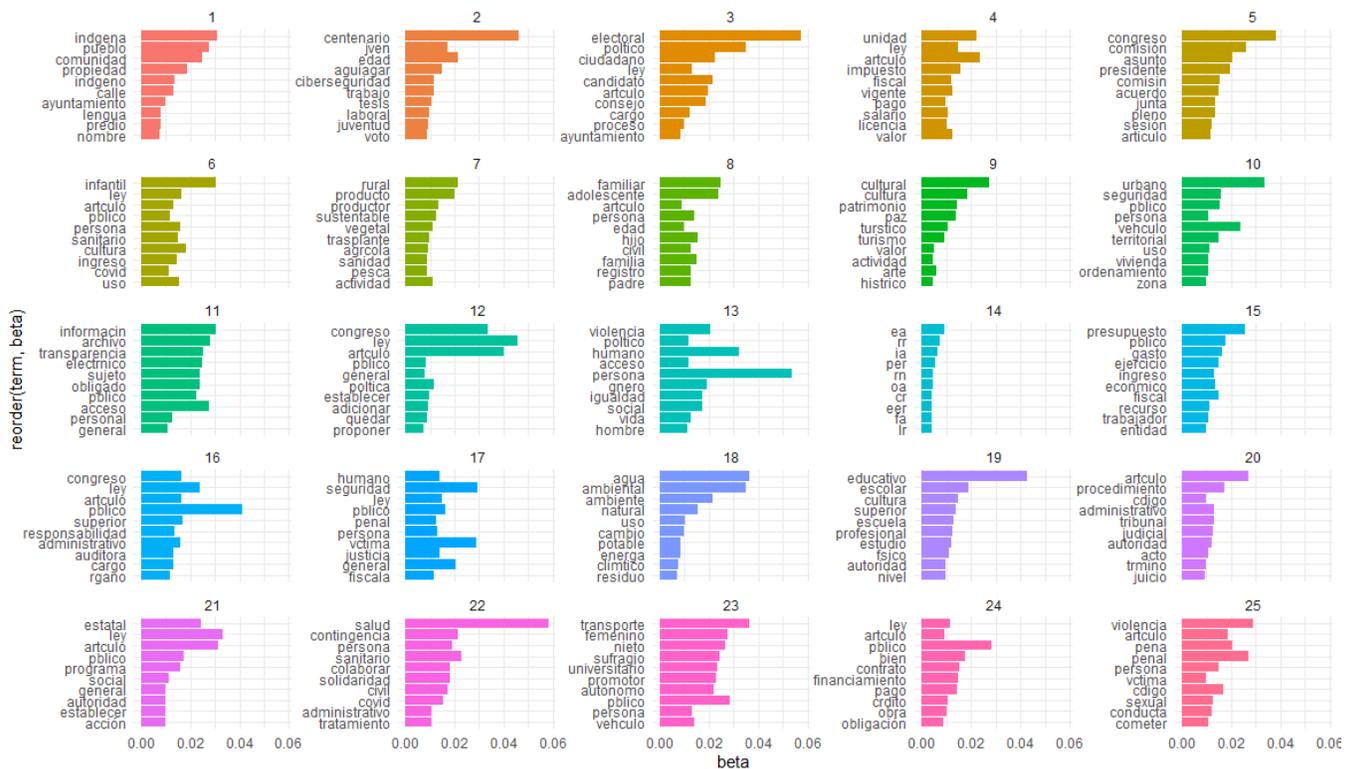


Figura 8: Tópicos y sus palabras claves

El análisis de tópicos en las propuestas legislativas del Congreso de San Luis Potosí (Tabla 8), revela una tendencia notable donde los temas administrativos y burocráticos dominan sobre aquellos de interés social. Esta observación encuentra respaldo en diversos estudios académicos

sobre la función y actividades de los cuerpos legislativos. Por ejemplo, el análisis de la estructura y funciones de las legislaturas estatales en EE.UU. destaca que una gran parte de su tiempo se dedica a la investigación, redacción y aprobación de legislaciones que a menudo tienen un carácter administrativo, como la elaboración de presupuestos y la supervisión de la rama ejecutiva. Estas actividades incluyen la gestión de comités, revisión de propuestas legislativas y coordinación con la agenda del gobernador, lo que puede limitar el tiempo y recursos disponibles para abordar problemas sociales directamente «Governors and State Legislatures» (2022). Además, estudios sobre la administración pública han señalado que los procesos legislativos están frecuentemente orientados hacia la creación y mantenimiento de estructuras administrativas eficaces. Esto se debe a la necesidad de gestionar de manera eficiente las funciones del gobierno y asegurar la implementación de políticas dentro del marco administrativo existente Sarker (2019).

Tabla 8: Proporción de tópicos administrativos frente a la suma de todos los demás tópicos

Año	Administrativos (%)	Otros-Tópicos (%)
2016	40.85711	59.14289
2017	38.50086	61.49914
2018	42.80298	57.19702
2019	38.83602	61.16398
2020	37.40853	62.59147
2021	37.99290	62.00710
2022	36.74962	63.25038

La Figura 9 muestra la proporción de tópicos administrativos versus la suma de los otros tópicos de las legislaturas del Congreso de San Luis Potosí entre 2016 y 2022. Observamos que en tópicos administrativos la proporción disminuye ligeramente desde 40.86% en 2016 hasta 36.75% en 2022. Y en “Otros-tópicos” la proporción aumenta de 59.14% en 2016 a 63.25% en 2022. Esto nos indica una tendencia hacia la disminución del enfoque en asuntos administrativos y un aumento en otros tipos de tópicos, mas enfocados en el interés social. Es así que dada la alta proporción de tópicos administrativos, decidimos omitirlos en los siguientes análisis para centrarnos en los tópicos de interés social. Esto reduce nuestro enfoque a 16 tópicos. A continuación vamos a analizar cómo ha cambiado su prioridad a lo largo de los años en las tres legislaturas.

Figura 10 y Figura 11 presentan los 16 tópicos restantes con sus proporciones reescaladas a lo largo de las tres legislaturas. Figura 10 muestra la evolución de todos los tópicos año por año, permitiendo observar las tendencias y cambios específicos en cada periodo legislativo. Figura 11 ilustra la distribución general de los tópicos a lo largo de los años, sin una separación anual, ofreciendo una visión global y continua de las proporciones temáticas.

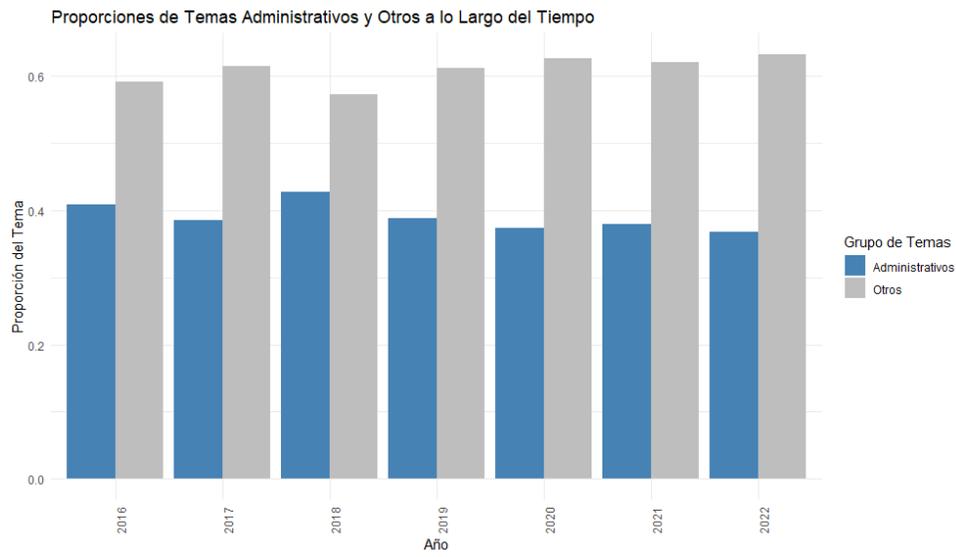


Figura 9: Tópicos Administrativos vs Otros a lo largo del tiempo

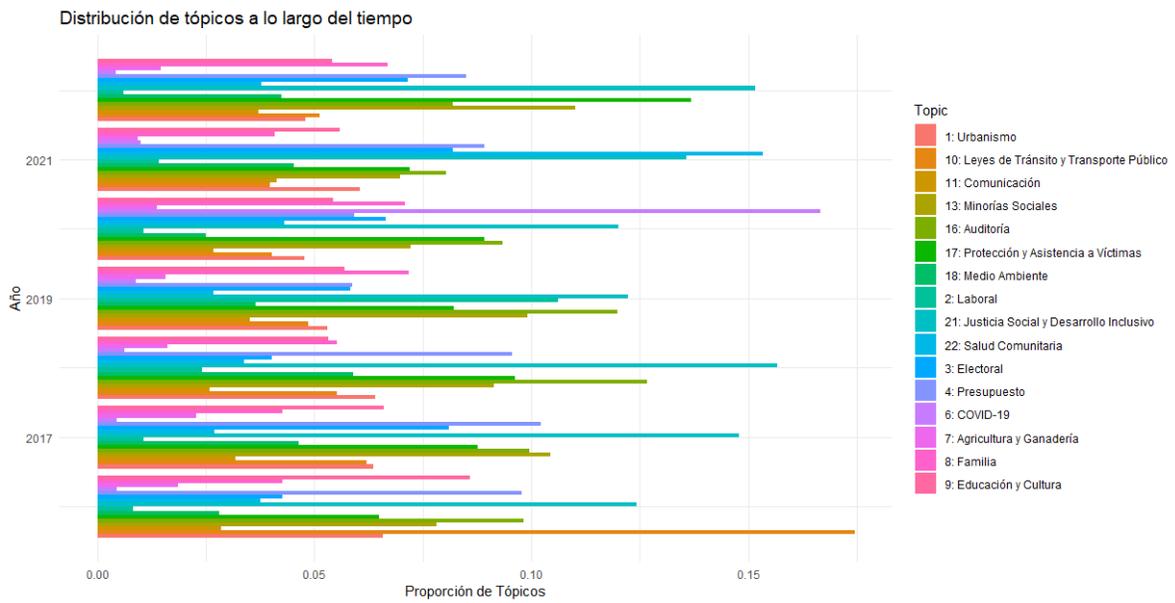


Figura 10: Proporción de Tópicos por Año

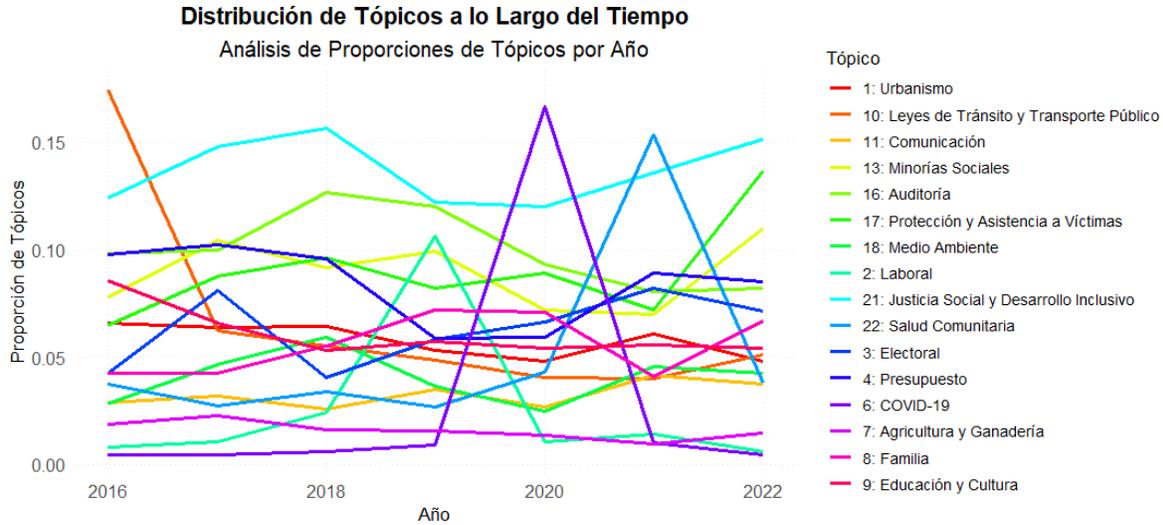


Figura 11: Evolución Temporal de la Distribución de Tópicos

6.3 Análisis Comparativo de Rankings y Proporciones

Dado que es complicado observar los cambios en la distribución de los tópicos a lo largo del tiempo, hemos desarrollado un ranking para identificar los tópicos que han experimentado las mayores modificaciones. Así definimos: El ranking del cambio de prioridades en los tópicos del congreso de San Luis Potosí, como un análisis que mide cómo han variado las prioridades de los temas abordados en el congreso a lo largo de diferentes periodos legislativos (ver Figura 12). Este ranking se basa en la proporción promedio de cada tópico en los documentos del congreso durante cada periodo específico, y cómo estas proporciones han cambiado entre los periodos⁷.

Durante los tres periodos legislativos del Congreso de San Luis Potosí, se observaron cambios significativos en los tópicos abordados y su relevancia, influenciados por diversos factores políticos, sociales y económicos. Este análisis nos permite visualizar cuáles tópicos han ganado o perdido importancia a lo largo del tiempo en el congreso, reflejando así las dinámicas y cambios en las prioridades legislativas.

⁷Consulte la Sección 7.4 para obtener la definición completa y el método detallado de cálculo del ranking en los diversos tópicos.

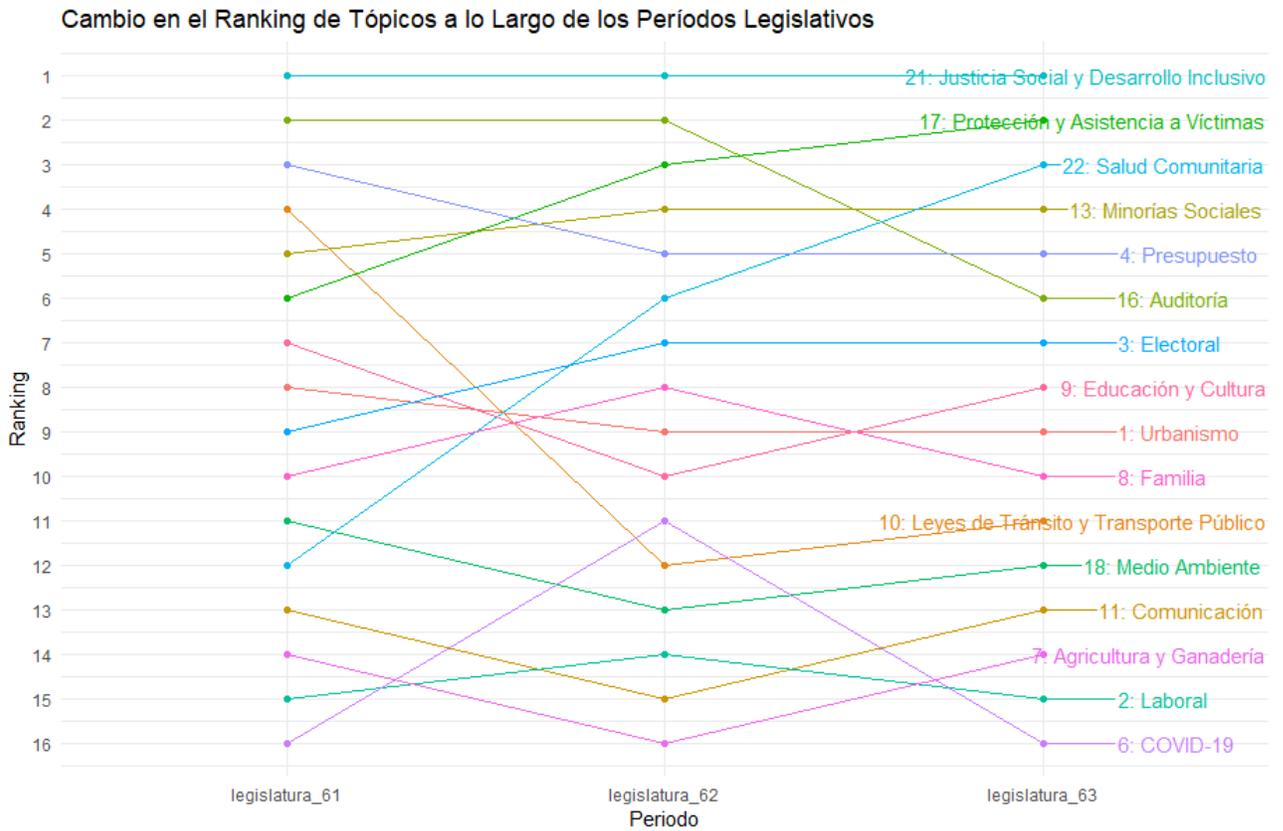


Figura 12: Cambio del Ranking en los Tópicos a lo largo de las 3 Legislaturas

Tabla 9: Cambio de Ranking de los Tópicos en las diferentes Legislaturas

Tópico	LXI	LXII	LXIII	Δ
6: Sanitario COVID	16	11	16	10
10: Leyes de Tránsito y Transporte Público	4	12	11	9
22: Salud Comunitaria	12	6	3	9
9: Educación y Cultura	7	10	8	5
17: Protección y Asistencia a Víctimas	6	3	2	4
16: Auditoría	2	2	6	4
8: Familia	10	8	10	4
11: Comunicación	13	15	13	4
7: Agricultura y Ganadería	14	16	14	4
18: Medio Ambiente	11	13	12	3
4: Presupuesto	3	5	5	2
3: Electoral	9	7	7	2
2: Laboral	15	14	15	2
13: Minorías Sociales	5	4	4	1
1: Urbanismo	8	9	9	1
21: Justicia Social y Desarrollo Inclusivo	1	1	1	0

6.4 Cambio de Ranking por Tópico y por Legislatura

El análisis de la Tabla 9 detalla los cambios en la relevancia de los 16 tópicos legislativos del Congreso de San Luis Potosí, ordenados de mayor a menor durante los periodos 2015-2018, 2018-2021 y 2021-2024. Se argumentan las razones detrás de estos cambios y se citan fuentes relevantes para respaldar la información.

6: Sanitario COVID - Cambio de Rango- 16 \rightarrow 11 \rightarrow 16

Este tópico vio un aumento en relevancia durante la pandemia, pero disminuyó después, reflejando la respuesta inmediata y la posterior adaptación a la nueva normalidad post-pandemia (Juanita (2021)).

10: Leyes de Tránsito y Transporte Público- Cambio de Rango: 4 \rightarrow 12 \rightarrow 11

La fluctuación puede estar asociada a cambios en la administración y las políticas de movilidad, así como a la respuesta a problemas urbanos emergentes y la implementación de nuevas infraestructuras de transporte (Mato, Daniel (2018)).

22: Salud Comunitaria - Cambio de Rango- 12 → 6 → 3

La salud comunitaria ha ganado importancia, especialmente en el contexto de la pandemia de COVID-19, que ha resaltado la necesidad de fortalecer los sistemas de salud pública y asegurar el bienestar comunitario (FID, SLP (2023)).

9: Educación y Cultura - Cambio de Rango- 7 → 10 → 8

La educación y cultura se mantienen como áreas de interés clave, especialmente con la implementación de nuevas reformas educativas y programas culturales que buscan preservar el patrimonio y fomentar la educación inclusiva (quadratin, slp (2023)).

17: Protección y Asistencia a Víctimas- Cambio de Rango: 6 → 3 → 2

La relevancia de este tópico ha aumentado, impulsada por un mayor enfoque en la justicia social y la respuesta a la violencia, así como a la implementación de políticas de apoyo a víctimas de delitos y desastres (razon, online (2024)).

16: Auditoría- Cambio de Rango: 2 → 2 → 6

El interés por la auditoría fue alto en los primeros dos periodos pero disminuyó ligeramente en el tercer período. Esto puede estar relacionado con la consolidación de sistemas de transparencia y fiscalización más robustos durante los primeros dos periodos (Congreso San Luis Potosí (s. f.)).

8: Familia- Cambio de Rango: 10 → 8 → 10

El interés en tópicos familiares ha fluctuado ligeramente, posiblemente influenciado por cambios en la composición demográfica y las necesidades sociales emergentes relacionadas con el bienestar familiar y la cohesión social (DIF, SLP (2023)).

11: Comunicación- Cambio de Rango-13 → 15 → 13

La comunicación sigue siendo crucial, reflejando la necesidad de transparencia, acceso a la información y la importancia de los medios en la sociedad (Congreso, SLP (2020)).

7: Agricultura y Ganadería- Cambio de Rango- 14 → 16 → 14

La agricultura y ganadería han mantenido su relevancia, destacando la importancia del sector primario en la economía regional y la necesidad de políticas de apoyo rural (Olivo (2023)).

18: Medio Ambiente- Cambio de Rango: 11 → 13 → 12

El medio ambiente ha visto una ligera variabilidad en su relevancia, reflejando la respuesta a desafíos ambientales crecientes y la implementación de políticas de sostenibilidad y conservación (Candidato, SLP (2023)).

4: Presupuesto- Cambio de Rango- 3 → 5 → 5

El presupuesto ha sido consistentemente relevante, destacando la importancia de la asignación de recursos para proyectos estatales y su adecuada distribución, especialmente en contextos de austeridad y reestructuración económica (Luis (2023)).

3: Electoral- Cambio de Rango- 9 → 7 → 7

Las reformas y discusiones electorales se han mantenido estables, evidenciando la importancia de mejorar la transparencia y la eficiencia en los procesos electorales para fortalecer la democracia local (quadratin (2023)).

2: Laboral- Cambio de Rango- 15 → 14 → 15

El interés en tópicos laborales ha sido constante, reflejando la importancia de las reformas laborales y la protección de los derechos de los trabajadores en un contexto de cambios económicos y sociales (Becerra (2024)).

13: Minorías Sociales- Cambio de Rango- 5 → 4 → 4

La protección de minorías sociales ha sido una prioridad constante, reflejando una creciente conciencia y legislación en derechos humanos y equidad social (Medrano, Maria (2024)).

1: Urbanismo- Cambio de Rango- 8 → 9 → 9

El urbanismo ha permanecido como un tópico importante, reflejando la necesidad de planificar y gestionar el crecimiento urbano de manera sostenible, especialmente en una era de rápida urbanización y cambio climático (razon, online 2023; Secretaría de Desarrollo Agrario, s. f.).

21: Justicia Social y Desarrollo Inclusivo- Constancia en la Relevancia

La justicia social y el desarrollo inclusivo han mantenido la máxima prioridad durante los tres periodos analizados, reflejando una constante preocupación por la equidad y el bienestar social, especialmente con la entrada de MORENA en 2018, que ha enfatizado políticas de inclusión y apoyo a sectores marginados (Mora, (2022)).

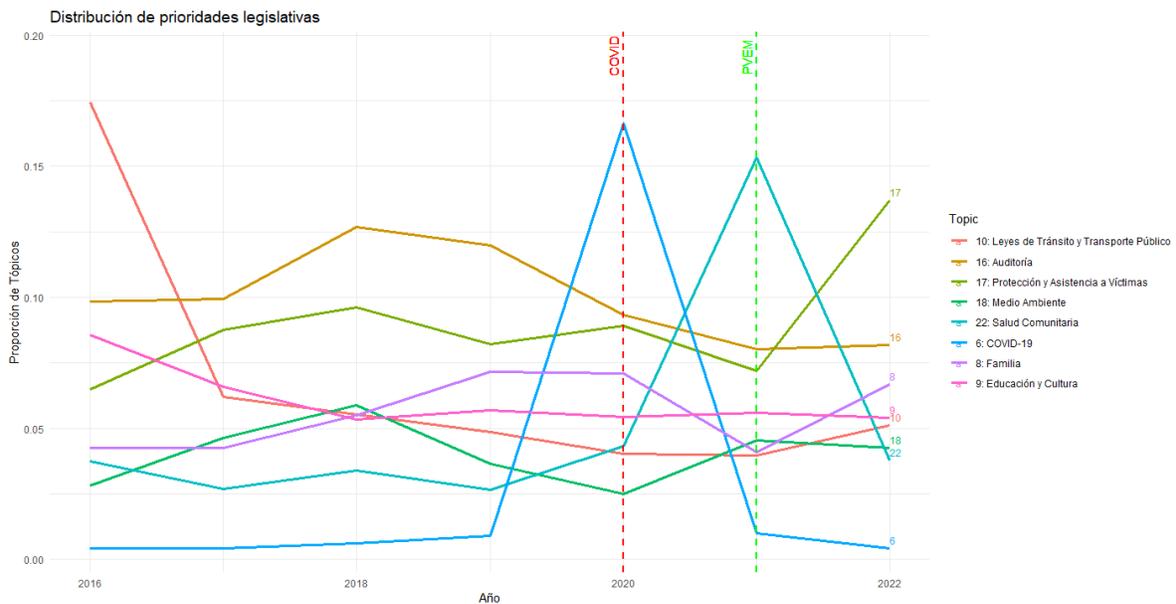


Figura 13: Tópicos con mayor variación

Un aspecto que llamó poderosamente nuestra atención es el cambio drástico en los tópicos, especialmente el tópico 6 relacionado con COVID-19 (ver Figura 13). En 2020, este tópico alcanzó su pico más alto debido a que era una prioridad nacional. Sin embargo, en 2021, cuando Ricardo Gallardo Cardona asumió funciones como gobernador de San Luis Potosí, el tópico de salud comunitaria disminuyó, mientras que el tópico sobre protección y asistencia a víctimas incrementó significativamente. Esto es curioso, ya que inicialmente pensamos que los temas de medio ambiente serían los más tratados, dado que el PVEM es un partido mayoritario. Sin embargo, al observar la Tabla 12, notamos que en esta legislatura, al dividir la proporción de tópicos por partidos mayoritarios, se observa lo contrario. Es más, el PRI y el PAN aportan mucho más que el PVEM, con proporciones de 24.88% y 8.05% respectivamente, mientras que el PVEM solo aporta un 0.49%.

Un fenómeno similar ocurre con el tópico 17, relacionado con la protección y asistencia a víctimas. Dentro de los partidos dominantes de la legislatura LXIII, el PAN aporta un 22.79%, el PRI un 8%, y el PVEM un 5.17%. En cuanto al tópico 22, salud comunitaria, el PAN contribuye con un 27.12%, el PRI con un 11.63% y el PVEM con un 1.69%. Para el tópico 6, relacionado con medidas sanitarias por COVID-19, las aportaciones son del 33.62% por parte del PAN, 6.10% por parte del PRI, y solo un 0.76% por parte del PVEM. Esto refleja una dinámica interesante en la que los partidos tradicionales aportan más a estos tópicos, a pesar de la presencia mayoritaria del PVEM y contar con un gobernador del mismo partido.

6.5 Prioridades Legislativas de los Partidos Dominantes

Para determinar qué partido político predominaba en cada legislatura, se investigó el número de legisladores por partido y, en base a esta proporción⁸, se identificaron los tres partidos más importantes de cada legislatura. Posteriormente, los tópicos extraídos del modelo LDA fueron filtrados según el partido político mayoritario de cada legislatura y se analizaron las proporciones de estos tópicos para evaluar el nivel de importancia asignado por cada partido. Este enfoque permitió observar cómo variaban las proporciones temáticas, obteniendo así un análisis detallado de las prioridades legislativas de los partidos dominantes.

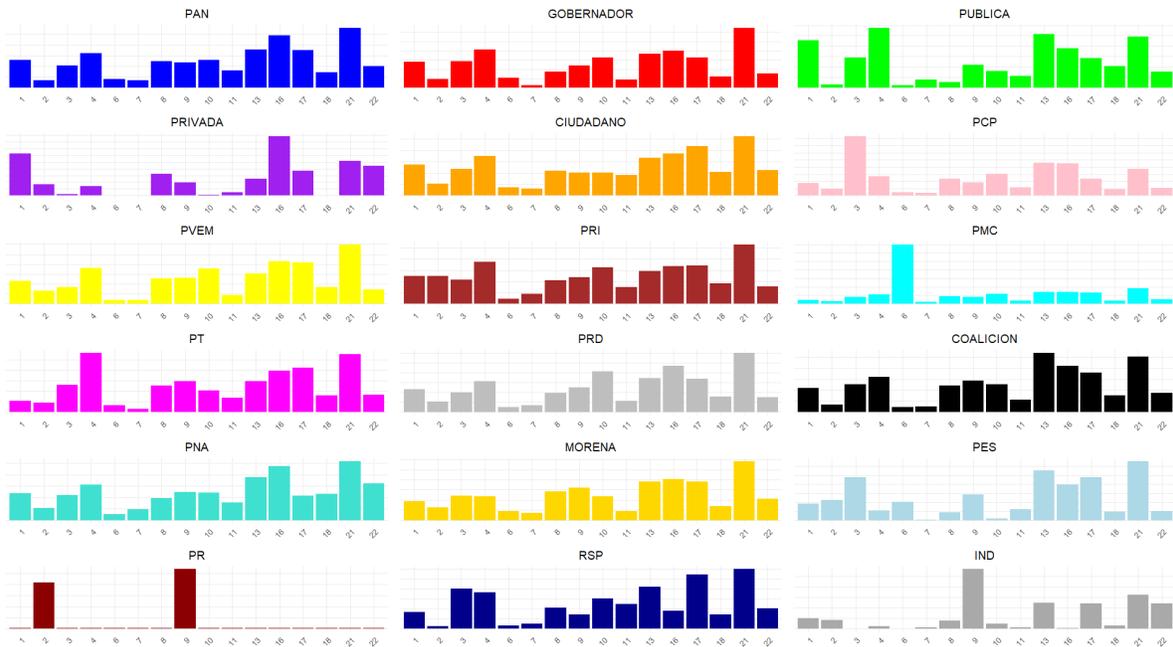


Figura 14: Distribución de Tópicos por Proponentes

⁸Consulte la Sección 7.1.

Tabla 10: Legislatura LXI (2015-2018)

Tópico	PRI (%)	PAN (%)	PRD (%)	OTROS (%)
1: Urbanismo	32.12	15.28	12.73	39.88
2: Laboral	32.96	11.41	17.50	38.14
3: Electoral	30.51	15.26	8.35	45.88
4: Presupuesto	31.11	16.55	10.17	42.18
6: Sanitario COVID	27.68	15.08	17.45	39.80
7: Agricultura y Ganadería	45.26	14.14	13.45	27.14
8: Familia	33.00	20.23	7.87	38.91
9: Educación y Cultura	25.96	16.99	11.64	45.41
10: Leyes de Tránsito y Transporte Público	29.27	14.88	14.90	40.96
11: Comunicación	32.34	17.97	10.54	39.15
13: Minorías Sociales	23.70	17.13	12.45	46.72
16: Auditoría	22.98	21.49	11.92	43.60
17: Protección y Asistencia a Víctimas	30.10	16.57	12.21	41.13
18: Medio Ambiente	30.64	14.75	11.56	43.05
21: Justicia Social y Desarrollo Inclusivo	27.97	16.12	13.75	42.17
22: Salud Comunitaria	30.75	13.70	13.56	41.99

Tabla 11: Legislatura LXII (2018-2021)

Tópico	PAN (%)	MORENA (%)	PRI (%)	OTROS (%)
1: Urbanismo	22.71	16.16	16.23	44.91
2: Laboral	9.33	14.56	41.77	34.34
3: Electoral	15.28	14.50	11.18	59.04
4: Presupuesto	14.60	13.33	17.34	54.73
6: Sanitario COVID	11.28	10.18	6.12	72.41
7: Agricultura y Ganadería	20.44	24.64	17.22	37.70
8: Familia	15.79	23.31	14.00	46.90
9: Educación y Cultura	16.60	27.00	15.42	40.99
10: Leyes de Tránsito y Transporte Público	19.62	24.01	17.98	38.39
11: Comunicación	22.41	13.59	20.59	43.41
13: Minorías Sociales	17.84	18.95	13.06	50.16
16: Auditoría	21.56	17.56	11.86	49.02

Tópico	PAN (%)	MORENA (%)	PRI (%)	OTROS (%)
17: Protección y Asistencia a Víctimas	16.02	18.98	18.92	46.09
18: Medio Ambiente	18.00	15.83	22.88	43.28
21: Justicia Social y Desarrollo Inclusivo	20.67	19.37	17.44	42.53
22: Salud Comunitaria	21.47	20.56	14.49	43.48

Tabla 12: Legislatura LXIII (2021-2024)

Tópico	PAN (%)	PVEM (%)	PRI (%)	OTROS (%)
1: Urbanismo	26.22	2.31	11.30	60.17
2: Laboral	14.02	8.31	25.43	52.25
3: Electoral	14.10	1.18	5.76	78.96
4: Presupuesto	21.59	1.59	10.52	66.30
6: Sanitario COVID	33.62	0.76	6.10	59.52
7: Agricultura y Ganadería	38.17	0.57	4.37	56.89
8: Familia	27.61	5.74	7.17	59.48
9: Educación y Cultura	17.83	3.77	10.84	67.56
10: Leyes de Tránsito y Transporte Público	19.22	6.64	6.38	67.75
11: Comunicación	29.37	0.18	15.70	54.75
13: Minorías Sociales	21.02	5.57	7.33	66.08
16: Auditoría	23.16	2.33	10.71	63.79
17: Protección y Asistencia a Víctimas	22.79	5.17	8.00	64.04
18: Medio Ambiente	24.88	0.49	8.05	66.58
21: Justicia Social y Desarrollo Inclusivo	24.05	5.11	9.42	61.42
22: Salud Comunitaria	27.12	1.69	11.63	59.57

Observaciones por Partido

PRI

- En la Legislatura LXI, el PRI se enfocó mayormente en “Agricultura y Ganadería” (45.26%) y “Laboral” (32.96%).
- En la Legislatura LXII, su mayor proporción fue en “Laboral” (41.77%) seguido por “Presupuesto” (17.34%).

- En la Legislatura LXIII, el PRI se enfocó en “Laboral” (25.43%) y “Presupuesto” (10.52%).

PAN

- En la Legislatura LXI, el PAN tuvo mayores proporciones en “Auditoría” (21.49%) y “Comunicación” (17.97%).
- En la Legislatura LXII, sus enfoques fueron “Urbanismo” (22.71%) y “Comunicación” (22.41%).
- En la Legislatura LXIII, los mayores enfoques fueron “Agricultura y Ganadería” (38.17%) y “Sanitario COVID” (33.62%).

PRD

- En la Legislatura LXI, el PRD se centró en “Laboral” (17.50%) y “Sanitario COVID” (17.45%).
- En la Legislatura LXII, sus enfoques más altos fueron “Laboral” (41.77%) y “Familia” (14.00%).
- En la Legislatura LXIII, no hay datos específicos ya que el PRD no figura entre los principales partidos.

OTROS

- En todas las legislaturas, los partidos clasificados como “OTROS” han tenido la mayor proporción en múltiples tópicos, con una presencia dominante en “Sanitario COVID” en la Legislatura LXII (72.41%) y “Electoral” en la Legislatura LXIII (78.96%).

Cambios Drásticos Observados

- **Sanitario COVID:** Hubo un cambio significativo en la proporción del tópico relacionado con COVID-19, con una disminución notable del enfoque de los partidos tradicionales y un aumento en los “OTROS” partidos.
- **Leyes de Tránsito y Transporte Público:** Este tópico vio una disminución en su ranking de importancia, moviéndose de la posición 4 en la Legislatura LXI a 12 y 11 en las legislaturas siguientes.
- **Salud Comunitaria:** Este tópico subió de la posición 12 en la Legislatura LXI a 6 y finalmente a 3 en la Legislatura LXIII, indicando un creciente interés en este tema.
- **Justicia Social y Desarrollo Inclusivo:** Este tópico mantuvo una posición constante en el ranking (1) en todas las legislaturas, mostrando su importancia continua.

El análisis revela que ciertos tópicos han visto cambios drásticos en su prioridad entre los partidos políticos a lo largo de las legislaturas. La proporción y el enfoque temático varían significativamente, con los partidos clasificados como “OTROS” consistentemente dominando varios tópicos clave. Este análisis proporciona una visión clara de cómo los intereses y prioridades de los partidos políticos evolucionan con el tiempo.

6.6 Análisis Jerárquico de Aglomeración

La aglomeración jerárquica se emplea para agrupar las propuestas legislativas en función de sus similitudes, con un enfoque especial en los partidos políticos. Esta técnica identifica clústeres de propuestas que reflejan enfoques semejantes entre distintos partidos, revelando coaliciones informales y alineamientos ideológicos. Visualizando estos resultados en dendrogramas, se pueden observar las relaciones jerárquicas entre las propuestas, mostrando cómo ciertos temas o enfoques están interrelacionados entre los distintos partidos. Este análisis facilita la identificación de grupos naturales de propuestas, descubriendo patrones en el comportamiento legislativo de los partidos políticos. Por ejemplo, se observa cómo ciertos partidos tienden a proponer legislación en temas específicos y cómo estas tendencias se alinean o divergen con las de otros partidos, proporcionando una visión clara de las dinámicas políticas y las posibles colaboraciones u oposiciones dentro del ámbito legislativo (Bunge y Judson 2005; Lima et al. 2023).

6.6.1 Dendograma

Un dendograma es una herramienta visual que representa jerárquicamente las relaciones de similitud entre los elementos, en este caso, los partidos políticos, mediante la agrupación en un árbol. Representa el resultado de un análisis de clustering jerárquico, que agrupa los datos secuencialmente, comenzando con cada dato como un grupo individual y luego combinándolos en grupos más grandes hasta formar un solo grupo. Los dendrogramas, construidos utilizando técnicas de agrupamiento jerárquico como el enlace simple, completo o promedio, muestran estas fusiones en forma de un árbol, donde la altura de las ramas indica la distancia o disimilitud entre los grupos. Este tipo de gráfico es útil para visualizar la estructura de los datos y determinar la cantidad óptima de clusters. (Hastie, Trevor, Tibshirani, Robert, y Friedman, Jerome 2009; Everitt et al. 2011).

6.6.2 Análisis Jerárquico y Visualización de Similitudes entre Proponentes

Para analizar las propuestas legislativas del Congreso de San Luis Potosí, se emplearon métodos de análisis estadístico y visualización de datos, enfocados en 18 proponentes que incluyen partidos políticos, el gobernador y sectores privados y públicos. Se calculó la proporción de tópicos para cada grupo, filtrando y agrupando los datos de proporciones tópicas normalizadas para cada partido. Esto permitió obtener la proporción de cada tópico en relación con el total de tópicos propuestos por cada proponente. Los datos resultantes se combinaron en una matriz de proporciones tópicas, donde las filas representaban a los proponentes y las columnas a los tópicos. Esta matriz fue escalada y se calcularon las distancias euclidianas entre los partidos. Posteriormente, se aplicó un análisis de clustering jerárquico con el método de enlace promedio para construir un dendograma, visualizando las similitudes y diferencias entre los

proponentes en términos de sus proporciones tópicas (M. Roberts, Stewart, y Tingley 2016; Grimmer 2010).

El análisis jerárquico permitió agrupar las propuestas legislativas según sus similitudes, destacando las similitudes entre partidos políticos y revelando coaliciones informales y alineamientos ideológicos. La visualización de los resultados en dendrogramas mostró cómo ciertos tópicos o enfoques están interrelacionados entre los distintos partidos, facilitando la identificación de grupos naturales de propuestas y descubriendo patrones en el comportamiento legislativo. Por ejemplo, se observó cómo ciertos partidos tienden a proponer legislación en tópicos específicos y cómo estas tendencias se alinean o divergen con las de otros partidos, proporcionando una visión clara de las dinámicas políticas y las posibles colaboraciones u oposiciones dentro del ámbito legislativo(Cha, Sung, s. f.; Dembele y Lo 2015).

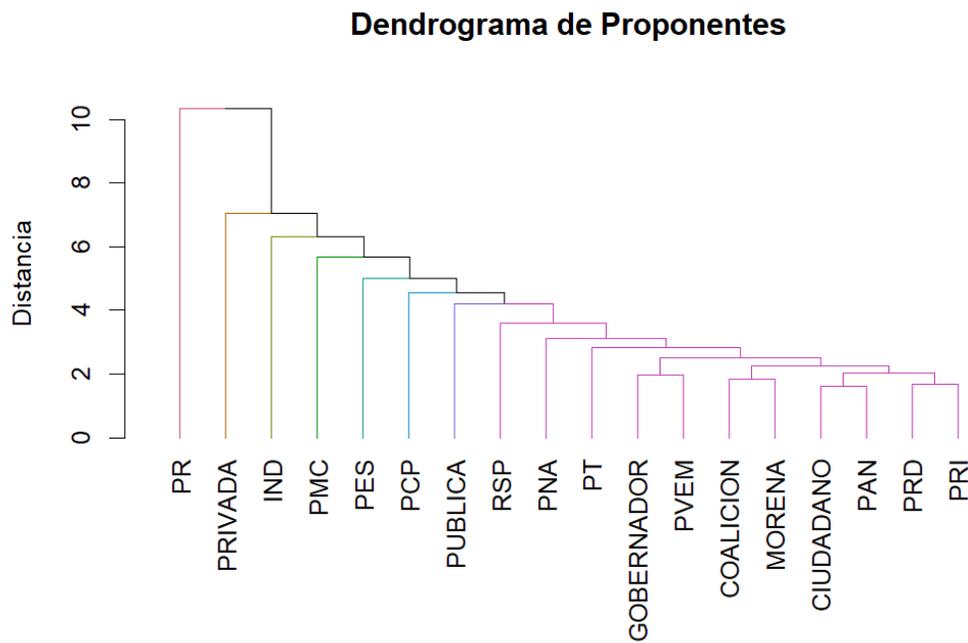


Figura 15: Dendrograma de Proporciones de Tópicos de los Proponentes en las Legislaturas

Aplicación en Análisis Político

El análisis de las proporciones de tópicos entre los partidos políticos, representado mediante el mapa de calor Figura 16 y el dendrograma Figura 15, revela patrones complejos de similitud y disimilitud que son cruciales para entender las dinámicas temáticas en la arena legislativa. A

continuación, se presenta un análisis detallado de estos patrones, resaltando las agrupaciones significativas y las diferencias notables entre los partidos.

Dendrograma de Proponentes

1. Análisis de Ramas y Clusters:

-*PR y PRIVADA*: Estos partidos se encuentran en ramas separadas y alejadas en el dendrograma, lo que reafirma su disimilitud observada en el mapa de calor. Esta separación implica enfoques legislativos radicalmente diferentes.

-*PVEM, Gobernador y MORENA*: Estos partidos están estrechamente agrupados en el dendrograma, lo que confirma la alta similitud temática observada en el mapa de calor. Esta proximidad sugiere una fuerte alineación en sus propuestas legislativas, posiblemente derivada de alianzas políticas o ideológicas.

-*RSP y otros partidos*: RSP se agrupa con PNA y PT, aunque no tan cercanamente como PVEM y MORENA, indicando una similitud temática moderada. Esto puede reflejar un enfoque temático compartido en ciertos aspectos, pero con diferencias suficientes para distinguirlos como grupos separados.

2. Patrones de Agrupamiento:

- Partidos Tradicionales: PRI, PRD y PAN tienden a agruparse juntos, mostrando similitudes en sus proporciones de tópicos. Esta agrupación puede reflejar enfoques temáticos comunes, posiblemente influenciados por su longevidad y experiencia en la política.
- Partidos Independientes y Menores: Partidos como IND y PRIVADA se agrupan más cerca de PMC y PES, sugiriendo similitudes temáticas, aunque mantienen diferencias claras con los partidos tradicionales. Esto puede indicar un enfoque temático más diverso o especializado.

Mapa de Calor

1. Interpretación de Colores y Distancias:

-Color Azul: Indica una distancia cero (como lo indica la diagonal en Figura 16), es decir, máxima similitud en las proporciones de tópicos entre los partidos comparados.

-Color Rojo: Representa una mayor distancia, sugiriendo una disimilitud significativa en las proporciones de tópicos.

2. Cluster Principales:

-*Gobernador y PVEM*: Estos dos proponentes muestran una casi perfecta similitud (color azul), lo que es coherente dado que ambos pertenecen al mismo partido. Esto refleja una alineación temática casi idéntica en sus propuestas legislativas.

-*MORENA y PVEM*: La similitud considerable entre MORENA y PVEM sugiere una convergencia temática, posiblemente debido a alianzas políticas o enfoques compartidos en sus agendas legislativas.

-*PR y PRIVADA*: La fuerte disimilitud (color rojo) entre PR y PRIVADA destaca diferencias marcadas en sus proporciones de tópicos, lo que puede reflejar estrategias legislativas y prioridades temáticas divergentes.

-*RSP y PR*: Similarmente, la disimilitud entre RSP y PR indica diferencias significativas en sus enfoques temáticos, posiblemente reflejando diferencias ideológicas o estratégicas.

Alineación Temática Fuerte: La alta similitud entre el Gobernador, PVEM y MORENA no solo refleja alianzas políticas actuales, sino también una coherencia en sus prioridades temáticas, lo que puede ser crucial para la implementación de políticas conjuntas.

Diversidad Temática: La disimilitud marcada entre PR y PRIVADA, así como entre RSP y PR, destaca la diversidad temática en el panorama legislativo. Esto puede ser indicativo de diferencias ideológicas profundas o de estrategias adaptativas para captar diferentes segmentos del electorado.

Estrategias y Prioridades: Las similitudes y diferencias en las proporciones de tópicos revelan estrategias y prioridades que no son evidentes a simple vista. Por ejemplo, la similitud temática entre MORENA y PVEM indica una clara coordinación en la formulación de políticas, mientras que la disimilitud entre PR y PRIVADA muestra una competencia directa en campos temáticos específicos.

Impacto en la Legislación: Estos patrones de similitud y disimilitud pueden tener un impacto significativo en la dinámica legislativa, afectando desde la formación de coaliciones hasta la negociación y aprobación de propuestas. Comprender estos patrones permite prever posibles alianzas y oposiciones dentro del cuerpo legislativo.

Este análisis de las proporciones de tópicos no solo proporciona una visión detallada de las dinámicas temáticas entre los partidos políticos, sino que también ofrece una base sólida para futuras investigaciones sobre el impacto de estas dinámicas en la gobernanza y la formulación de políticas. La identificación de patrones de similitud y disimilitud ayudará a los investigadores a entender mejor cómo las prioridades temáticas influyen en la política legislativa y en la interacción entre diferentes actores políticos.

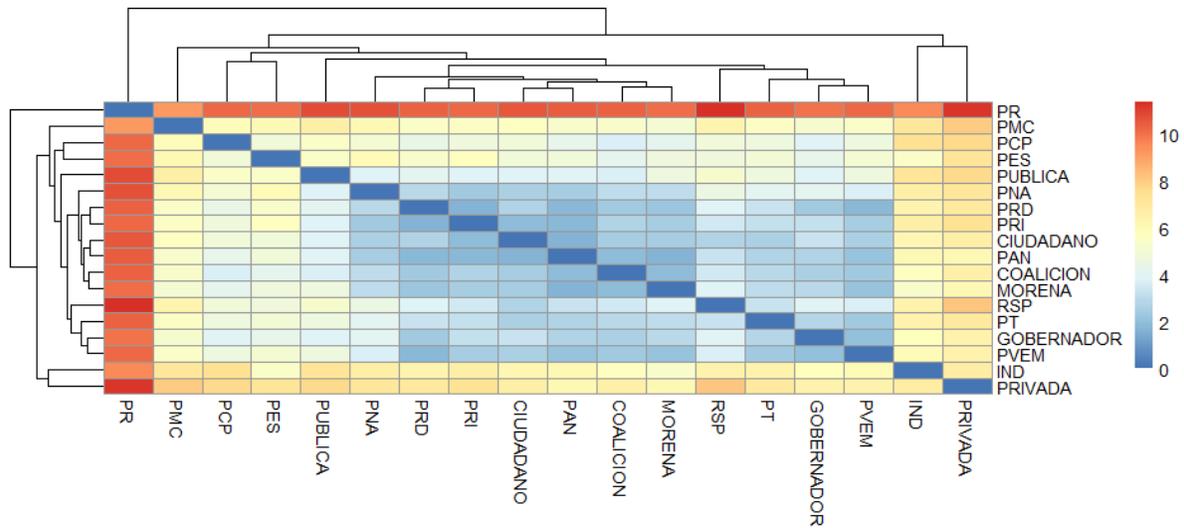


Figura 16: Mapa de Calor de Proporciones de Tópicos de Proponentes

6.7 Aportes de la Tesis

La presente tesis ofrece una serie de contribuciones significativas al análisis legislativo en el Congreso de San Luis Potosí durante el periodo 2016-2022. Este estudio se destaca por su enfoque innovador en la digitalización y procesamiento automático de documentos legislativos, permitiendo un análisis exhaustivo y detallado de las propuestas presentadas en este periodo. Además, proporciona una base empírica sólida y una metodología replicable que otros investigadores pueden utilizar para analizar dinámicas legislativas en diferentes contextos, facilitando la identificación de patrones temáticos y la formulación de políticas más informadas.

1. Digitalización y Procesamiento Automático

El primer gran aporte de esta tesis es la digitalización completa de más de 3,600 documentos legislativos mediante técnicas avanzadas de web scraping y reconocimiento óptico de caracteres (OCR). Este proceso permitió transformar documentos en formato PDF escaneado a texto editable, facilitando así el análisis posterior. La limpieza y filtrado de texto incluyeron la eliminación de palabras sin sentido y la tokenización de los documentos, mejorando la calidad y precisión del análisis.

2. Creación e Incorporación de Diccionarios

Se desarrollaron e incorporaron diversos diccionarios específicos para la eliminación de stopwords y palabras irrelevantes. Estos diccionarios fueron esenciales para la limpieza y preparación del texto, asegurando que solo las palabras y términos significativos fueran considerados en el análisis. Este paso fue crucial para optimizar el rendimiento de los modelos de análisis de texto utilizados.

3. Identificación de Tópicos Abordados

Utilizando el modelo Latent Dirichlet Allocation (LDA), se lograron identificar y categorizar 17 tópicos principales en los documentos legislativos analizados. Este modelo permitió descubrir la estructura temática subyacente de las propuestas, revelando patrones y tendencias importantes en la actividad legislativa del Congreso de San Luis Potosí.

4. Prioridades Legislativas por Partido

El análisis de los tópicos identificados permitió determinar las prioridades legislativas de los diferentes partidos políticos representados en el congreso. Se observó una variabilidad significativa en las prioridades de los partidos, destacando similitudes y diferencias notables entre ellos. Por ejemplo, se identificaron afinidades temáticas entre el gobernador y el PVEM, y diferencias significativas entre partidos como PR y PRIVADA, y RSP y PR.

5. Aporte a la Literatura: Tópicos Administrativos y Burocráticos

Una contribución destacada de esta tesis es la cuantificación de la proporción de temas administrativos y burocráticos en comparación con otros tópicos de interés social. Se observó que, aunque los temas administrativos dominan las propuestas legislativas, su proporción ha disminuido ligeramente del 40.86% en 2016 al 36.75% en 2022. Este hallazgo proporciona una base empírica para futuras investigaciones sobre la eficiencia y enfoque de los cuerpos legislativos.

6. Análisis de Similitud entre Proponentes

El estudio incluyó un análisis jerárquico de aglomeración para visualizar las similitudes y diferencias temáticas entre los proponentes de las legislaturas. Este análisis reveló patrones de afinidad entre ciertos partidos y proponentes, proporcionando una comprensión más profunda de las dinámicas políticas y legislativas en el Congreso de San Luis Potosí.

7 Apéndices

7.1 Partidos predominantes

Legislatura LXI

Tabla 13: Número de diputados por partido político de la legislatura LXI Leon, Oscar (2015)

Abrev.	Partido político	Legisladores	%
PRI	Partido Revolucionario Institucional	8	29.63
PAN	Partido Acción Nacional	7	25.93
PRD	Partido de la Revolución Democrática	4	14.81
PVEM	Partido Verde Ecologista de México	2	7.41
PNA	Nueva Alianza	2	7.41
PT	Partido del Trabajo	1	3.70
PMC	Movimiento Ciudadano	1	3.70
MORENA	Movimiento Regeneración Nacional	1	3.70
PCP	Partido Conciencia Popular	1	3.70

Legislatura LXII

Tabla 14: Número de diputados por partido político de la legislatura LXII Congreso, SLP (2017)

Abrev.	Partido político	Legisladores	%
PAN	Partido Acción Nacional	6	22.22
MORENA	Movimiento Regeneración Nacional	6	22.22
PRI	Partido Revolucionario Institucional	5	18.52
PT	Partido del Trabajo	2	7.41
PVEM	Partido Verde Ecologista de México	2	7.41
PRD	Partido de la Revolución Democrática	2	7.41
PNA	Nueva Alianza	1	3.70
PMC	Movimiento Ciudadano	1	3.70
PES	Partido Encuentro Social	1	3.70
PCP	Partido Conciencia Popular	1	3.70

Legislatura LXIII

Tabla 15: Número de diputados por partido político de la legislatura LXIII Manuel, Carreras (2023)

Abrev.	Partido político	Legisladores	%
PAN	Partido Acción Nacional	6	22.22
PVEM	Partido Verde Ecologista de México	6	22.22
PRI	Partido Revolucionario Institucional	4	14.81
MORENA	Movimiento Regeneración Nacional	4	14.81
PT	Partido del Trabajo	3	11.11
PMC	Movimiento Ciudadano	1	3.70
RSP	Redes Sociales Progresistas	1	3.70
PNASLP	Nueva Alianza San Luis Potosí	1	3.70
PCP	Partido Conciencia Popular	1	3.70

7.2 Métricas para Determinar el Número de Tópicos en LDA

Métricas evaluadas

Griffiths y Steyvers (2004) esta métrica se basa en la probabilidad de los datos dados los parámetros del modelo, utilizando una técnica de muestreo de Gibbs colapsado. Se selecciona el número de tópicos (k) que maximiza la probabilidad de los datos.

Ventajas: - Metodología probabilística robusta. - Adecuada para conjuntos de datos grandes y complejos. - Produce modelos interpretables.

Desventajas: - Requiere un cálculo intensivo. - Sensible a la elección de hiperparámetros.

Cao et al. (2009) propone una métrica basada en la densidad que mide la distancia media entre distribuciones de tópicos. El valor de (k) que minimiza esta métrica se considera óptimo.

Ventajas: - Metodología intuitiva y directa. - Fácil de implementar. - Buena rendimiento en diversos tipos de datos.

Desventajas: - Puede ser menos precisa en conjuntos de datos con alta dimensionalidad. - No siempre produce resultados consistentes con otros métodos.

Arun et al. (2010) utiliza la divergencia KL entre distribuciones de tópicos y documentos para determinar (k). La métrica considera la relación entre la matriz de palabras por tópicos y la matriz de documentos por tópicos.

Ventajas: - Método basado en principios teóricos sólidos. - Permite una interpretación clara de los resultados. - Específicamente diseñada para LDA.

Desventajas: - Computacionalmente intensivo. - Requiere ajuste fino de parámetros.

Deveaud, SanJuan, y Bellot (2014) introduce una métrica basada en la coherencia de tópicos, la cual mide la similitud promedio de las palabras más probables dentro de los tópicos. El (k) óptimo maximiza esta coherencia.

Ventajas: - Produce tópicos más interpretables y coherentes. - Menos sensible a la elección de hiperparámetros. - Fácil de entender e implementar.

Desventajas: - Puede no funcionar bien en conjuntos de datos extremadamente grandes. - En algunos casos, puede ser necesario combinar con otras métricas para una selección óptima.

7.3 Descripción del Proceso LDA

Modelado de Tópicos

En este capítulo, se introduce la técnica de modelado de tópicos, una herramienta poderosa en el campo del procesamiento del lenguaje natural. Esta técnica busca construir tópicos o temas en base a las distribuciones de palabras en un conjunto de documentos. A lo largo de este ejercicio, se explorarán:

- Cómo pre-procesar texto para su posterior análisis.
- Cómo construir vectores de documentos por términos.
- Cómo modelar tópicos usando LDA (Latent Dirichlet Allocation).
- Cómo interpretar los tópicos, leyendo los resultados junto con un análisis cualitativo.

Preprocesamiento de Texto

El primer paso en el procesamiento del lenguaje natural es la carga y preprocesamiento del texto. Las tareas de preprocesamiento incluyen:

- Análisis morfosintáctico para determinar los componentes de la oración.
- Lematización para reducir las palabras a sus formas básicas.
- Eliminación de palabras comunes y sin sentido, quedándose solo con las más significativas.

Vectorizado del Texto

Para construir las tablas a partir de las palabras del documento, se utiliza la función `count()` y `cast_dtm` de la librería `tidytext`, que permiten crear una matriz de términos-documentos (DTM).

Modelado de Tópicos con LDA

El modelo LDA se usa para construir los tópicos. Los resultados se exportan en los formatos necesarios para su análisis utilizando la función `tidy`.

Interpretación del Modelo

El etiquetado de tópicos es similar a la codificación cualitativa. Se listan los términos que más contribuyen a cada tópico, y se escriben estas etiquetas en un array de nombres de tópicos para su uso en futuros gráficos.

Consideraciones Adicionales

Se discuten decisiones sobre la organización y agrupación de tópicos, como la posible eliminación de tópicos irrelevantes o la agrupación de tópicos similares.

En este proyecto se ha llevado a cabo un Análisis de Asignación Latente (LDA) sobre datos legislativos para identificar los tópicos subyacentes. A continuación, se describen en detalle los pasos realizados a lo largo del código:

Paso 0: Instalación de Paquetes

Primero, se instalaron varios paquetes necesarios para realizar el análisis de texto y visualizaciones. Estos incluyen `tm`, `topicmodels`, `reshape2`, `ggplot2`, `wordcloud`, `pals`, `SnowballC`, `lda`, `ldatuning`, `kableExtra`, `flextable`, `remotes`, y `ggthemes`. Estos paquetes proporcionan las herramientas necesarias para manejar, procesar y visualizar los datos textuales.

Paso 1: Cargar los Paquetes

Se cargaron las librerías necesarias, incluyendo `lubridate`, `furrr`, `readxl`, `tidyverse`, `tidytext`, `stm`, `quanteda`, `qdapRegex`, `dplyr`, `tm`, `stopwords`, `udpipe`, `topicmodels`, `wordcloud`, `ldatuning`, `ggplot2`, `pals`, `RColorBrewer`, `ggrepel`, `scales`, `grDevices` y `ggthemes`. Estas librerías ofrecen funcionalidades para el manejo de fechas, paralelización, lectura de archivos Excel, manipulación de datos, análisis de texto, modelado de tópicos, y creación de visualizaciones.

Paso 2: Cargar y Limpiar los Datos

Se cargaron los datos legislativos y se filtraron para incluir solo aquellos registros entre el 1 de enero de 2016 y el 31 de diciembre de 2022. Posteriormente, se verificaron los valores faltantes y se eliminó cualquier registro incompleto. Se utilizó un diccionario personalizado para eliminar palabras sin sentido. Luego, se realizó una limpieza adicional del texto, eliminando números, puntuación y stop words utilizando varias fuentes de stop words en español (`stopwords-iso`, `snowball`, `nlTK`), así como las palabras del diccionario personalizado.

Paso 3: Lematización

Se descargó y cargó un modelo de lenguaje en español utilizando `udpipe`. El texto se anotó utilizando este modelo, y solo se conservaron las lemas que eran adjetivos, verbos o sustantivos, eliminando cualquier stop word adicional.

Paso 4: Preparación de la Matriz de Términos-Documentos (DTM)

Se creó una matriz de términos-documentos (DTM) contando las palabras por documento y convirtiendo estos conteos en una estructura adecuada para el modelado de tópicos. Se eliminaron las filas vacías de la DTM, ya que el modelo LDA no puede manejar documentos vacíos.

Paso 5: Búsqueda del Mejor Número de Tópicos (K)

Se evaluaron diferentes valores de K (número de tópicos) utilizando varias métricas (Griffiths2004, CaoJuan2009, Arun2010, Deveaud2014) para determinar el mejor número de tópicos. Se identificó el mejor valor de K y se seleccionó para el entrenamiento del modelo LDA.

Paso 6: Entrenamiento del Modelo LDA

Se entrenó un modelo LDA utilizando el mejor valor de K encontrado. Se extrajeron las matrices beta (distribución de palabras por tópicos) y gamma (distribución de tópicos por documentos).

Paso 7: Evaluación y Visualización de Resultados

Se analizaron y visualizaron los resultados del modelo LDA. Esto incluyó:

- Graficar las proporciones de los tópicos a lo largo del tiempo.
- Crear wordclouds para los términos más relevantes de cada tópico.
- Evaluar los cambios en los rankings de los tópicos entre diferentes periodos legislativos.
- Analizar la distribución promedio de los tópicos por partido político.

Paso 8: Clustering Jerárquico

Se realizó un análisis de clustering jerárquico para agrupar los partidos políticos en función de sus distribuciones de tópicos. Se generó un dendrograma para visualizar estas agrupaciones.

Paso 9: Otros Análisis y Visualizaciones

Se contaron el total de caracteres y palabras en los textos. Se realizaron visualizaciones adicionales para mostrar las proporciones de los tópicos más interesantes y relevantes.

Este flujo de trabajo proporciona un análisis detallado y estructurado de los datos legislativos utilizando técnicas de modelado de tópicos. Los resultados ofrecen una visión profunda de los tópicos predominantes y su evolución a lo largo del tiempo, así como una comprensión de cómo estos tópicos varían entre los diferentes partidos políticos.

7.4 Cálculo del Cambio de Ranking de los Tópicos Legislativos

Para calcular el cambio en el ranking de los tópicos legislativos entre las legislaturas, se utilizó la diferencia absoluta entre los rankings consecutivos. Este método es comúnmente utilizado en el análisis de datos y series temporales para medir cambios y variaciones (Montgomery y Runger 2010; Śpiewanowski y Talavera 2021). En particular, el uso de la diferencia absoluta para evaluar el cambio en rankings está bien documentado en la literatura de (Kendall 1938; Berenson, Mark et al. 2020).

En esta sección, se describe el proceso utilizado para calcular el cambio total en el ranking de los tópicos legislativos a lo largo de las legislaturas LXI, LXII y LXIII del Congreso de San Luis Potosí.

7.4.1 Cálculo del Cambio Total en el Ranking

Para cada tópico T_i , se calculó el cambio absoluto en el ranking entre cada par de períodos legislativos consecutivos usando la siguiente fórmula:

$$\Delta r_{i,j} = |r_{i,j+1} - r_{i,j}|$$

Donde:

- $r_{i,j}$ es el ranking del tópico T_i en el período j .
- $\Delta r_{i,j}$ es el cambio absoluto en el ranking del tópico T_i entre el período j y el período $j + 1$.

El cambio total en el ranking para el tópico T_i a lo largo de todos los períodos se calculó sumando los cambios absolutos entre cada par de períodos consecutivos:

$$\text{Cambio Total}(T_i) = \sum_{j=1}^{n-1} \Delta r_{i,j} = \sum_{j=1}^{n-1} |r_{i,j+1} - r_{i,j}|$$

7.4.1.1 Ejemplos de Cálculo

Ejemplo 1: Tópico 6 (Sanitario COVID)

Para el tópico 6:

- Legislatura 61: $r_{6,61} = 16$
- Legislatura 62: $r_{6,62} = 11$
- Legislatura 63: $r_{6,63} = 16$

El cambio absoluto en el ranking entre las legislaturas 61 y 62 es:

$$\Delta r_{6,61-62} = |r_{6,62} - r_{6,61}| = |11 - 16| = 5$$

El cambio absoluto en el ranking entre las legislaturas 62 y 63 es:

$$\Delta r_{6,62-63} = |r_{6,63} - r_{6,62}| = |16 - 11| = 5$$

El cambio total en el ranking para el t3pico 6 es:

$$\text{Cambio Total}(T_6) = \Delta r_{6,61-62} + \Delta r_{6,62-63} = 5 + 5 = 10$$

Ejemplo 2: T3pico 22 (Salud Comunitaria)

Para el t3pico 22:

- Legislatura 61: $r_{22,61} = 12$
- Legislatura 62: $r_{22,62} = 6$
- Legislatura 63: $r_{22,63} = 3$

El cambio absoluto en el ranking entre las legislaturas 61 y 62 es:

$$\Delta r_{22,61-62} = |r_{22,62} - r_{22,61}| = |6 - 12| = 6$$

El cambio absoluto en el ranking entre las legislaturas 62 y 63 es:

$$\Delta r_{22,62-63} = |r_{22,63} - r_{22,62}| = |3 - 6| = 3$$

El cambio total en el ranking para el t3pico 22 es:

$$\text{Cambio Total}(T_{22}) = \Delta r_{22,61-62} + \Delta r_{22,62-63} = 6 + 3 = 9$$

Ejemplo 3: T3pico 10 (Leyes de Tr3nsito y Transporte P3blico)

Para el t3pico 10:

- Legislatura 61: $r_{10,61} = 4$
- Legislatura 62: $r_{10,62} = 12$
- Legislatura 63: $r_{10,63} = 11$

El cambio absoluto en el ranking entre las legislaturas 61 y 62 es:

$$\Delta r_{10,61-62} = |r_{10,62} - r_{10,61}| = |12 - 4| = 8$$

El cambio absoluto en el ranking entre las legislaturas 62 y 63 es:

$$\Delta r_{10,62-63} = |r_{10,63} - r_{10,62}| = |11 - 12| = 1$$

El cambio total en el ranking para el t3pico 10 es:

$$\text{Cambio Total}(T_{10}) = \Delta r_{10,61-62} + \Delta r_{10,62-63} = 8 + 1 = 9$$

7.5 Ejemplo computacional

El artículo de Steyvers y Griffiths (2004) fue el primero en utilizar el muestreo de Gibbs para estimar las variables latentes de LDA, siguiendo a Steyvers y Griffiths tenemos:

- z_i es el tópico asignado al i -ésimo token en toda la colección;
- d_i es el documento que contiene el i -ésimo token;
- w_i es el tipo de palabra del i -ésimo token;
- \mathbf{z}_{-i} es el conjunto de asignaciones de tópicos de todos los demás tokens;
- \cdot es cualquier información restante, como los hiperparámetros α y β :

$$P(z_i = j | \mathbf{z}_{-i}, w_i, d_i, \cdot) = \frac{C_{w_i j}^{WT} + \beta}{\sum_{w=1}^W C_{w j}^{WT} + W\beta} \frac{C_{d_i j}^{DT} + \alpha}{\sum_{t=1}^T C_{d_i t}^{DT} + T\alpha} \quad (10)$$

donde \mathbf{C}^{WT} y \mathbf{C}^{DT} son matrices de conteos (palabra-tópico y documento-tópico)

Estimaciones de ϕ y θ

$$\phi_{ij} = \frac{C_{ij}^{WT} + \beta}{\sum_{k=1}^W C_{kj}^{WT} + W\beta} \quad \theta_{dj} = \frac{C_{dj}^{DT} + \alpha}{\sum_{k=1}^T C_{dk}^{DT} + T\alpha}$$

Usando las matrices de conteo como antes, donde ϕ_{ij} es la probabilidad del tipo de palabra i para el tópico j , y θ_{dj} es la proporción del tópico j en el documento d

Ejemplo en Código

Para este ejemplo usaremos un corpus de 8 documentos, que es extraído de las legislaturas en estudio.

```
propuestas<- c('genera confusión en la práctica',
              'ello es así en virtud que al estar abordando el tema del acusado',
              'sin más separación que una coma',
              'se ocupa del diverso referente a los alimentos',
              'luego entonces la confusión que en ocasiones se propicia',
              'agua san luis potosi legislatura',
              'acusado legislador al que actual',
              'la actual redacción del precepto legal cuya modificación se plantea'
              )
```

Generamos la lista de documentos:

```
docs <- strsplit(propuestas, split=' ', perl = TRUE)
str(docs)
```

List of 8

```
$ : chr [1:5] "genera" "confusión" "en" "la" ...
$ : chr [1:13] "ello" "es" "así" "en" ...
$ : chr [1:6] "sin" "más" "separación" "que" ...
$ : chr [1:8] "se" "ocupa" "del" "diverso" ...
$ : chr [1:9] "luego" "entonces" "la" "confusión" ...
$ : chr [1:5] "agua" "san" "luis" "potosi" ...
$ : chr [1:5] "acusado" "legislador" "al" "que" ...
$ : chr [1:10] "la" "actual" "redacción" "del" ...
```

Palabras únicas para conformar el vocabulario *V*

```
vocab <- unique(unlist(docs))
str(vocab)
```

```
chr [1:46] "genera" "confusión" "en" "la" "práctica" "ello" "es" "así" ...
```

Abstrayendo las palabras para tener un problema numérico. Reemplazamos las palabras en documentos con WordIDs, y reemplazamos las palabras en documentos con identificadores de palabras:

```
for(i in 1:length(docs))
  docs[[i]] <- match(docs[[i]],vocab)
head(vocab)
```

```
[1] "genera"      "confusión" "en"          "la"          "práctica" "ello"
```

```
docs[[2]]
```

```
[1] 6 7 8 3 9 10 11 12 13 14 15 16 17
```

```
docs[[4]]
```

```
[1] 23 24 16 25 26 27 28 29
```

Definiendo el número de tópicos a buscar

```
K <- 3
```

Inicializar matrices de conteo w_t : matriz palabra-tema

```
wt <- matrix(0,K,length(vocab))
colnames(wt) <- vocab
```

t_d : lista de asignación de tópicos

```
ta <- lapply(docs, function(x) rep( 0, length(x)))
names(ta) <- paste0( "doc", 1:length(docs) )
```

d_t : conteos que corresponden al número de palabras asignadas a cada tópico para cada documento

```
dt <- matrix(0, length(docs),K)
set.seed(1234)

for(d in 1:length(docs))
  {# Asignar un tópico aleatoriamente a cada palabra w
  for(w in 1:length( docs[[d]]))
    {
      ta[[d]][w] <- sample(1:K,1)
      # Extraer el índice del tópico, el id de la palabra y actualizar la celda correspondiente
      # en la matriz de conteo palabra-tema
      ti <- ta[[d]][w]
      wi <- docs[[d]][w]
      wt[ti,wi] <- wt[ti,wi] + 1
    }
  # Contar palabras en el documento d asignadas a cada tópico t

  for(t in 1:K)
    dt[d,t] <- sum(ta[[d]] == t)
  }
```

Conteo de cada palabra asignada a cada tópico

Lista de asignación de tópicos

```
print(ta$doc2)
```

```
[1] 1 2 2 3 2 2 2 3 2 2 2 1
```

```
print(wt[, 1:4])
```

```
      genera confusión en la
[1,]      0          0  1  0
[2,]      1          2  0  1
[3,]      0          0  2  2
```

```
print(dt)
```

```
      [,1] [,2] [,3]
[1,]     2     2     1
[2,]     2     9     2
[3,]     2     1     3
[4,]     1     2     5
[5,]     2     5     2
[6,]     3     1     1
[7,]     2     1     2
[8,]     2     4     4
```

Muestreo de Gibbs ejemplo en una iteración

Fijamos los Hiperparámetros

```
alpha <- 1
beta <- 1
```

Temas iniciales asignados a la primera palabra del primer documento y su correspondiente id de palabra

```
t0 <- ta[[1]][1]
wid <- docs[[1]][1]
```

z_{-i} significa que no incluimos el token w en nuestras matrices de conteo palabra-tema y documento-tema al muestrear para el token w , solo dejamos las asignaciones de temas de todos los otros tokens para el documento 1.

```
dt[1,t0] <- dt[1,t0]-1
wt[t0,wid] <- wt[t0,wid]-1
```

Calcular lado izquierdo y derecho de la Ecuación 10

```

izquierda <- (wt[,wid] + beta) / (rowSums(wt)+length(vocab)*beta)
derecha <- (dt[1,] + alpha) / (sum(dt[1,]) + K*alpha)

```

Extraer un nuevo t3pico para la primera palabra en el primer documento

```

t1 <- sample(1:K,1, prob = izquierda * derecha)
t1

```

```
[1] 1
```

Actualizar d_t y w_t con el t3pico reci3n asignado

```

ta[[1]][1] <- t1
dt[1,t1] <- dt[1,t1] + 1
wt[t1,wid] <- wt[t1,wid] + 1

```

Muestreo de Gibbs usando la librer3a topicmodels

Definimos los par3metros

```

K <- 3
alpha <- 1
beta <- .001
iteraciones <- 1000
source("./LDA_functions.R")
set.seed(4321)
lda1 <- LDA1( docs = docs,
              vocab = vocab, K = K, alpha = alpha, eta = beta,
              iterations = iteraciones )

```

C3lculo de la Probabilidad posterior

Probabilidad del t3pico de cada palabra:

```
phi <- (lda1$wt + beta) / (rowSums(lda1$wt) + length(vocab) * beta)
```

Probabilidad de t3pico de cada documento:

```
theta <- (lda1$dt + alpha) / (rowSums(lda1$dt) + K * alpha)
```

Topico asignado a cada documento, calculamos el que tiene la mayor probabilidad

```
topic <- apply(theta, 1, which.max)
```

Palabras posibles bajo cada t3pico

Ordenamos la probabilidad y obtenemos el n3mero especificado por el usuario

```
Terms <- function(phi,n)
{
  term <- matrix( 0, n, K )
  for( p in 1:nrow(phi) )
    term[ , p ] <- names( sort( phi[ p, ], decreasing = TRUE )[1:n] )
  return(term)
}
```

```
term <- Terms(phi = phi, n = 3)
```

```
list(texto_original = propuestas[topic == 1], palabras = term[,1])
```

```
$texto_original
```

```
[1] "ello es as3 en virtud que al estar abordando el tema del acusado"
[2] "luego entonces la confusi3n que en ocasiones se propicia"
[3] "acusado legislador al que actual"
```

```
$palabras
```

```
[1] "que" "en" "se"
```

```
list(texto_original = propuestas[topic == 2], palabras = term[,2])
```

```
$texto_original
```

```
[1] "se ocupa del diverso referente a los alimentos"
```

```
$palabras
```

```
[1] "del" "ello" "abordando"
```

Comparamos con el paquete oficial de topicmodels

```
library(tm)
library(topicmodels)
# @burning : n3mero de iteraciones de Gibbs omitidas al principio
# @thin : n3mero de iteraciones de Gibbs omitidas entre medio
```

```
docs1 <- Corpus( VectorSource(propuestas))
dtm <- DocumentTermMatrix(docs1)
lda <- LDA(dtm, k = 3, method = "Gibbs",
           control = list( seed = 1234, burnin = 500, thin = 100, iter = 4000))

list(texto_original = propuestas[topics(lda) == 1], palabras = terms(lda,3)[,1])
```

```
$texto_original
```

```
[1] "genera confusión en la práctica"
[2] "luego entonces la confusión que en ocasiones se propicia"
```

```
$palabras
```

```
[1] "del"          "confusión" "acusado"
```

```
list(texto_original = propuestas[topics(lda) == 2], palabras = terms(lda,3)[,2])
```

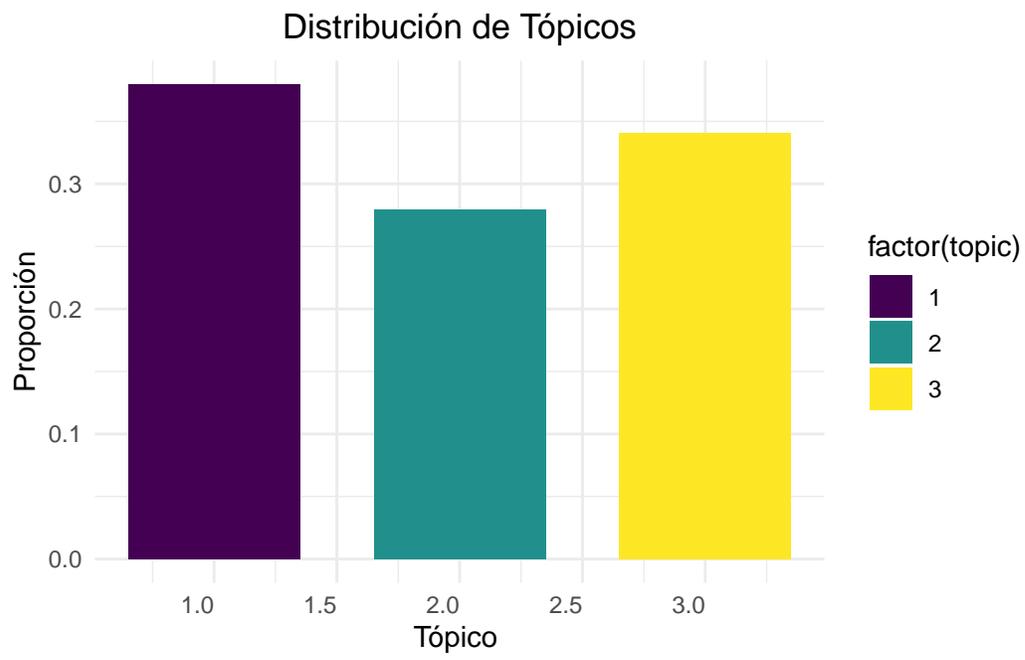
```
$texto_original
```

```
[1] "ello es así en virtud que al estar abordando el tema del acusado"
[2] "acusado legislador al que actual"
[3] "la actual redacción del precepto legal cuya modificación se plantea"
```

```
$palabras
```

```
[1] "actual"      "práctica"  "abordando"
```

Graficando la proporción de tópicos:



8 Conclusiones

Esta investigación ha ofrecido un análisis exhaustivo y profundo de las propuestas legislativas en el Congreso de San Luis Potosí durante las legislaturas LXI, LXII y LXIII, abarcando desde el 2016 hasta el 2022. Utilizando técnicas avanzadas de procesamiento de lenguaje natural, específicamente el modelo Latent Dirichlet Allocation (LDA), se lograron identificar y categorizar 17 tópicos en más de 3,600 documentos legislativos. Este enfoque ha revelado hallazgos significativos que destacan la evolución de las prioridades legislativas y las dinámicas políticas en San Luis Potosí.

Predominio de Temas Administrativos y Cambio hacia Tópicos Sociales

Un descubrimiento central es la predominancia de temas administrativos y burocráticos, que inicialmente constituyeron el 40.86% de las propuestas en 2016, disminuyendo a un 36.75% en 2022. Este cambio indica una tendencia creciente hacia temas de interés social, reflejando una respuesta adaptativa del Congreso a las necesidades emergentes de la comunidad. Este desplazamiento hacia temas sociales destaca una mayor sensibilidad y reacción a los desafíos contemporáneos, como la salud comunitaria y la justicia social.

Impacto de la Pandemia de COVID-19

La pandemia de COVID-19 ha dejado una huella significativa en las prioridades legislativas. Hubo un notable aumento temporal en las propuestas relacionadas con la salud comunitaria durante el periodo más crítico de la pandemia. Sin embargo, post-pandemia, la relevancia de estos tópicos disminuyó, mostrando una adaptación a la nueva normalidad y a otras prioridades emergentes.

Evolución de la Justicia Social y Desarrollo Inclusivo

Los temas de justicia social y desarrollo inclusivo se han mantenido consistentemente importantes a lo largo de las tres legislaturas. Este constante enfoque en la equidad y el bienestar social resalta el compromiso del Congreso con la protección y promoción de los derechos humanos y la inclusión de las minorías sociales. La relevancia de estos tópicos subraya un reconocimiento de las necesidades de las comunidades más vulnerables y una búsqueda continua de justicia y desarrollo integral.

Análisis de Género en las Propuestas Legislativas

El análisis de género ha revelado una distribución desigual en la participación legislativa entre hombres y mujeres. Partidos como el PRI y MORENA destacan por una alta actividad legislativa femenina, con 351 y 301 propuestas presentadas por mujeres respectivamente. Sin embargo, otros partidos muestran una notable disparidad de género, como el PAN, donde los hombres presentaron el doble de propuestas que las mujeres (427 frente a 218). Este análisis subraya la necesidad de promover una mayor equidad de género en la participación legislativa, abriendo la puerta para futuras investigaciones y políticas que fomenten la inclusión y representación equitativa.

Visualización de Similitudes y Diferencias entre Partidos Políticos

El análisis jerárquico de aglomeración y la visualización de similitudes temáticas entre los partidos políticos han revelado patrones interesantes de afinidad y diferencias. Se ha identificado una notable similitud entre las propuestas del gobernador y PVEM, mientras que hay diferencias significativas entre partidos como PR y PRIVADA, y RSP y PR. Estos patrones ofrecen una visión clara de las alianzas y oposiciones dentro del Congreso, proporcionando una base sólida para comprender las dinámicas políticas y estratégicas.

Conclusión Final y Futuras Implicaciones

Este estudio no solo proporciona una comprensión integral de las dinámicas legislativas en San Luis Potosí, sino que también establece una base empírica robusta para futuras investigaciones y la formulación de políticas más informadas. La aplicación del análisis cuantitativo en un ámbito tradicionalmente cualitativo ha demostrado ser una herramienta eficaz para medir la actividad política y legislativa, ofreciendo nuevas perspectivas y enfoques para la mejora continua de la gobernanza y la respuesta a las necesidades sociales.

El análisis de textos a través de algoritmos computacionales provee al investigador en ciencias sociales y económicas con herramientas para extraer contenido y analizar estructuras de discurso difícilmente accesibles a través de métodos tradicionales. Algoritmos de análisis accesibles permiten no solo abordar grandes volúmenes de texto, sino que proveen herramientas de análisis estructural como la identificación automática de tópicos, análisis de sentimiento en textos o medios sociales e inclusive la exploración de autoría de obras históricas o artísticas.

La inclusión del análisis de género resalta la importancia de fomentar la participación equitativa de mujeres en la legislatura, un área que merece más atención y acción en investigaciones futuras. Este estudio sirve como una guía para futuros legisladores y académicos en la identificación y análisis de las tendencias y prioridades legislativas, promoviendo una política más inclusiva y equitativa para todos los ciudadanos de San Luis Potosí.

Bibliografía

- Argüelles, Elia. 2017. «Instituciones políticas y políticas públicas, el rol del Congreso del Estado de San Luis Potosí en la política de Seguridad Pública (2006-2015).» Tesis doctoral, SLP. <https://colsan.repositorioinstitucional.mx/jspui/handle/1013/501>.
- Arun, R., V. Suresh, C. E. Veni Madhavan, y M. N. Narasimha Murthy. 2010. «On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations». En, editado por Mohammed J. Zaki, Jeffrey Xu Yu, B. Ravindran, y Vikram Pudi, 391-402. Lecture Notes En Computer Science. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-13657-3_43.
- Becerra, Bertha. 2024. «Necesario incorporar a la formalidad a 32.5 millones de trabajadores: especialista». <https://www.elsoldesanluis.com.mx/finanzas/necesario-incorporar-a-la-formalidad-a-32.5-millones-de-trabajadores-especialista-11871350.html>.
- Bela A., Frigyik, Amol, Kapila, y Maya R, Gupta. 2010. «Introduction to the Dirichlet Distribution and Related Processes». Seattle, Washington. <https://vannevar.ece.uw.edu/techsite/papers/documents/UWEEETR-2010-0006.pdf>.
- Benoit, Kenneth, Kohei Watanabe, Haiyan Wang, Paul Nulty, Adam Obeng, Stefan Müller, y Akitaka Matsuo. 2018. «quanteda: An R package for the quantitative analysis of textual data» 3: 774. <https://doi.org/10.21105/joss.00774>.
- Benoit, Kenneth, Kohei Watanabe, Haiyan Wang, Patrick O. Perry, Benjamin Lauderdale, Johannes Gruber, y William Lowe. 2023. «quanteda.textmodels: Scaling Models and Classifiers for Textual Data». <https://CRAN.R-project.org/package=quanteda.textmodels>.
- Berenson, Mark, Levine, David, Szabat, Kathryn, y Stephan, David. 2020. *Basic Business Statistics: Concepts and Applications*. 5th ed. Pearson. <https://www.pearson.com/en-us/subject-catalog/p/basic-business-statistics-concepts-and-applications/P200000006245/9780137400119>.
- Blei, D. M. 2012. «Probabilistic topic models». *Commun. ACM* 55: 7784.
- Blei, D. M., A. Y. Ng, y M. I. Jordan. 2003. «Latent Dirichlet allocation». *J. Mach. Learn Res.* 3: 9931022.
- Blei, David M., y Michael I. Jordan. 2006. «Variational inference for Dirichlet process mixtures». *Bayesian Analysis* 1 (1): 121-43. <https://doi.org/10.1214/06-BA104>.
- Bunge, John A., y Dean H. Judson. 2005. «Data Mining». En, editado por Kimberly Kempf-Leonard, 617-24. New York: Elsevier. <https://doi.org/10.1016/B0-12-369398-5/00159-6>.
- Candidato, SLP. 2023. «SE ENTREGARON 200 MIL ARBOLITOS PARA LA SIERRA DE SAN MIGUELITO - El Candidato MX». <https://elcandidato.mx/2023/12/01/se-entregar-on-200-mil-arbolitos-para-la-sierra-de-san-miguelito/>.
- Cao, Juan, Tian Xia, Jintao Li, Yongdong Zhang, y Sheng Tang. 2009. «A density-based method for adaptive LDA model selection». *Neurocomputing, Advances en Machine Learning y Computational Intelligence*, 72 (7): 1775-81. <https://doi.org/10.1016/j.neucom.2008.06.011>.
- Cha, Sung. s. f. «Comprehensive Survey on Distance/Similarity Measures Between Probability Density Functions». *International Journal of Mathematical Models and Methods in Applied Sciences*.

- Chang, I-Cheng, Tai-Kuei Yu, Yu-Jie Chang, y Tai-Yi Yu. 2021. «Applying Text Mining, Clustering Analysis, and Latent Dirichlet Allocation Techniques for Topic Classification of Environmental Education Journals». *Sustainability* 13 (19): 10856. <https://doi.org/10.3390/su131910856>.
- Chang, Jonathan. 2015. «lda: Collapsed Gibbs Sampling Methods for Topic Models». <https://CRAN.R-project.org/package=lda>.
- Congreso San Luis Potosí. 2023. «Legislaturas Congreso de San Luis Potosí.» <https://congresosanluis.gob.mx/trabajo/trabajo-legislativo/iniciativas?page=1>.
- . s. f. «LXI Legislatura». https://es.wikipedia.org/wiki/Anexo:LXI_Legislatura_del_Congreso_del_Estado_de_San_Luis_Potos%C3%AD.
- Congreso, SLP. 2017. «LXII CANDIDATOS ELECTOS PROCESO ELECTORAL 2017-2018.» https://ceepacslp.org.mx/ceepac/uploads2/files/ELECTOS%20diputados%20M_R_%20y%20R_P_%202018.pdf.
- . 2020. «Transparencia | Congreso de San Luis Potosi». <https://congresosanluis.gob.mx/trabajo/transparencia>.
- . 2023. «Iniciativas | Congreso de San Luis Potosi». <https://congresosanluis.gob.mx/trabajo/trabajo-legislativo/iniciativas?page=1>.
- Curran, Ben, Kyle Higham, Elisenda Ortiz, y Demival Vasques Filho. 2018. «Look Who's Talking: Two-Mode Networks as Representations of a Topic Model of New Zealand Parliamentary Speeches». *PLOS ONE* 13 (6): e0199072. <https://doi.org/10.1371/journal.pone.0199072>.
- Dembele, Soumaila, y Gane Samb Lo. 2015. «Probabilistic, Statistical and Algorithmic Aspects of the Similarity of Texts and Application to Gospels Comparison». *Journal of Data Analysis and Information Processing* 3 (4): 112-27. <https://doi.org/10.4236/jdaip.2015.34012>.
- Deveaud, Romain, Eric SanJuan, y Patrice Bellot. 2014. «Accurate and Effective Latent Concept Modeling for Ad Hoc Information Retrieval». *Document Numerique* 17 (1): 61-84. <https://www.cairn.info/revue-document-numerique-2014-1-page-61.htm>.
- DIF, SLP. 2023. «San Luis Potosí despega en desarrollo social y económico | Revista Punto de Vista». <https://revistapuntodevista.com.mx/nflash/san-luis-potosi-despega-en-desarrollo-social-y-economico/843483/>.
- Everitt, Brian S., Sabine Landau, Morven Leese, y Daniel Stahl. 2011. *Cluster Analysis*. 1.^a ed. Wiley Series En Probability y Statistics. Wiley. <https://doi.org/10.1002/9780470977811>.
- FID, SLP. 2023. «Mejor calidad de vida con apoyo a la salud, impulsa patrimonio de la Beneficencia Pública Estatal | Revista Punto de Vista». <https://revistapuntodevista.com.mx/nflash/mejor-calidad-de-vida-con-apoyo-a-la-salud-impulsa-patrimonio-de-la-beneficencia-publica-estatal/822443/>.
- Geman, Stuart, y Donald Geman. 1984. «Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images». *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-6 (6): 721-41. <https://doi.org/10.1109/TPAMI.1984.4767596>.
- «Governors and State Legislatures». 2022. [https://socialsci.libretexts.org/Bookshelves/Political_Science_and_Civics/American_Government_3e_\(OpenStax\)/14%3A_State_and_Local_Government/14.04%3A_Governors_and_State_Legislatures](https://socialsci.libretexts.org/Bookshelves/Political_Science_and_Civics/American_Government_3e_(OpenStax)/14%3A_State_and_Local_Government/14.04%3A_Governors_and_State_Legislatures).
- Griffiths, Thomas L., y Mark Steyvers. 2004. «Finding scientific topics». *Proceedings of the*

- National Academy of Sciences* 101 (suppl_1): 5228-35. <https://doi.org/10.1073/pnas.0307752101>.
- Griffiths, Thomas, y Steyvers, Mark. 2007. «Finding Scientific Topics». En. <https://doi.org/10.1073/pnas.0307752101>.
- Grimmer, Justin. 2010. «A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases». *Political Analysis* 18 (1): 1-35. <https://doi.org/10.1093/pan/mpp034>.
- Grün, Bettina, y Kurt Hornik. 2023b. «topicmodels: Topic Models». <https://CRAN.R-project.org/package=topicmodels>.
- . 2023a. «topicmodels: Topic Models». <https://CRAN.R-project.org/package=topicmodels>.
- Hastie, Trevor, Tibshirani, Robert, y Friedman, Jerome. 2009. *The Elements of Statistical Learning*. Springer Series En Statistics. Springer. <https://link.springer.com/book/10.1007/978-0-387-84858-7>.
- Heinrich, Gregor. 2009. «Parameter Estimation for Text Analysis». Germany. <https://www.arbylon.net/publications/text-est2.pdf>.
- Jelodar, Hamed, Yongli Wang, Chi Yuan, y Xia Feng. 2017. «Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey», noviembre.
- Juanita, Olivo. 2021. «En SLP circulan 34 variantes del Covid-19; cada día más personas requieren ser intubadas: Salud». <https://www.elsoldesanluis.com.mx/local/en-slp-circulan-34-variantes-del-covid-19-cada-dia-mas-personas-requieren-ser-intubadas-salud-7039356.html>.
- Kendall, M. G. 1938. «A New Measure of Rank Correlation». *Biometrika* 30 (1/2): 81-93. <https://doi.org/10.2307/2332226>.
- Laver, Michael, Kenneth Benoit, y John Garry. 2003. «Extracting Policy Positions from Political Texts Using Words as Data». *American Political Science Review* 97 (02). <https://doi.org/10.1017/S0003055403000698>.
- Leon, Oscar. 2015. «LXI Consejo Estatal electoral y de Partición Ciudadana». *Periódico Oficial del Estado libre y soberano de San Luis Potosí*, 4.
- Lima, Willian P. C., Lucas C. Marques, Laura S. Assis, y Douglas O. Cardoso. 2023. «An Analysis of Political Parties Cohesion Based on Congressional Speeches». En, editado por Jiří Mikyška, Clélia De Mulatier, Maciej Paszynski, Valeria V. Krzhizhanovskaya, Jack J. Dongarra, y Peter M. A. Sloot, 10475:105-19. Cham: Springer Nature Switzerland. https://link.springer.com/10.1007/978-3-031-36024-4_8.
- Luis, Bianca Acuña. 2023. «Presupuesto Estatal SLP». <https://www.elsoldesanluis.com.mx/local/anuncia-sgg-integracion-del-presupuesto-de-egresos-2023-por-dependencias-estatales-8867132.html>.
- Luque, Carolina, Rubriche, Juan, Galvis, Jhon, y Sosa, Juan. 2021. «Modelamiento de tópicos para identificar patrones en la investigación científica del Covid-19». <https://revistas.usantotomas.edu.co/index.php/estadistica/article/view/7705/7237>.
- Manuel, Carreras. 2023. «LXIII Plan de San Luis». *Periodico Oficial del Estado*, 19.
- Mark Steyvers, Tom Griffiths. 2007. «Probabilistic Topic Models». En. Psychology Press.
- Mato, Daniel. 2018. *Educación superior, diversidad cultural e interculturalidad en América*

- Latina*. UNESCO. <https://unesdoc.unesco.org/ark:/48223/pf0000372649>.
- Medrano, Maria. 2024. «protección». <https://www.astrolabio.com.mx/mecanismo-de-proteccion-se-queda-corto-en-san-luis-potosi/>.
- Minka, Thomas, y John Lafferty. 2002. «Expectation-propagation for the generative aspect model». En, 352359. UAI'02. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.
- Montgomery, Douglas C., y George C. Runger. 2010. *Applied Statistics and Probability for Engineers*. John Wiley & Sons.
- Mora, Luis. 2022. «SLP Turismo Social e Inclusivo». <https://www.elsoldesanluis.com.mx/local/avanza-slp-en-turismo-social-e-inclusivo-8916332.html>.
- Nikita, Murzintcev. 2020. «ldatuning: Tuning of the Latent Dirichlet Allocation Models Parameters». <https://CRAN.R-project.org/package=ldatuning>.
- Odden, Tor Ole B., Alessandro Marin, y John L. Rudolph. 2021. «How Has *Science Education* Changed over the Last 100 Years? An Analysis Using Natural Language Processing». *Science Education* 105 (4): 653-80. <https://doi.org/10.1002/sce.21623>.
- Olivo, Juana. 2023. «Sequía en Villa de Arriaga; “es casi imposible ser agricultor o ganadero”: alcalde». <https://www.elsoldesanluis.com.mx/local/soledad/sequia-en-villa-de-arriaga-es-casi-imposible-ser-agricultor-o-ganadero-alcalde-10958254.html>.
- Ooms, Jeroen. 2022. «tesseract: Open Source OCR Engine». <https://CRAN.R-project.org/package=tesseract>.
- Ponweiser, Martin. 2012. «Latent Dirichlet Allocation in R». Tesis doctoral, Institute for Statistics; Mathematics No. 2.
- quadratin, slp. 2023. «Confirma Ceepac nueva distritación electoral en San Luis Potosí». <https://sanluispotosi.quadratin.com.mx/san-luis-potosi/confirma-ceepac-nueva-distritacion-electoral-en-san-luis-potosi/>.
- quadratin, slp. 2023. «2023, año lleno de actividades educativas y culturales en la capital SLP». <https://sanluispotosi.quadratin.com.mx/principal/2023-ano-lleno-de-actividades-educativas-y-culturales-en-la-capital-slp/>.
- razon, online. 2023. «San Luis Potosí se suma a esfuerzos para frenar cambio climático tras reunión de Conago». <https://www.razon.com.mx/estados/san-luis-potosi-suma-esfuerzos-frenar-cambio-climatico-reunion-conago-522069>.
- . 2024. «Con apoyo estatal, San Luis Potosí logra en 2023 menor incidencia de delitos en 6 años». <https://www.razon.com.mx/estados/apoyo-estatal-san-luis-potosi-logra-2023-menor-incidencia-delitos-6-anos-561240>.
- Roberts, Margaret E., Brandon M. Stewart, y Edoardo M. Airoldi. 2016. «A Model of Text for Experimentation in the Social Sciences». *Journal of the American Statistical Association* 111 (515): 988-1003. <https://doi.org/10.1080/01621459.2016.1141684>.
- Roberts, Margaret, Brandon Stewart, y Dustin Tingley. 2016. «Navigating the Local Modes of Big Data: The Case of Topic Models». *Computational Social Science: Discovery and Prediction*. <https://scholar.harvard.edu/dtingley/publications/navigating-local-modes-big-data-case-topic-models>.
- Sarker, Md Nazirul Islam. 2019. «Public Administration as an Academic Discipline and Social Science». En, 1-7. Springer, Cham. https://doi.org/10.1007/978-3-319-31816-5_3856-1.
- Secretaría de Desarrollo Agrario, Sedatu. s. f. «Presenta Sedatu estrategia de planeación te-

territorial para la Zona Metropolitana de San Luis Potosí». <http://www.gob.mx/sedatu/prensa/presenta-sedatu-estrategia-de-planeacion-territorial-para-la-zona-metropolitana-de-san-luis-potosi>.

Śpiewanowski, Piotr, y Oleksandr Talavera. 2021. «Journal Rankings and Publication Strategy». *Scientometrics* 126 (4): 3227-42. <https://doi.org/10.1007/s11192-021-03891-5>.

Steyvers, Mark, y Tom Griffiths. 2004. «Probabilistic Topic Models».

Wickham, Hadley. 2016. «ggplot2: Elegant Graphics for Data Analysis». <https://ggplot2.tidyverse.org>.

