CENTRO DE INVESTIGACIÓN EN

CIENCIAS DE INFORMACIÓN GEOESPACIAL, A.C.

CentroGeo

Centro Público de Investigación SECIHTI

GENERACIÓN AUTOMATIZADA DE TÓPICOS EN LA PLATAFORMA DE ACCESO A LA INFORMACIÓN PÚBLICA GUBERNAMENTAL (2003-2020)

TESIS

Que para obtener el grado de

Doctor en Ciencias de Información Geoespacial

Presenta

Hermelando Cruz Pérez

Director de Tesis

Dr. Alejandro Molina Villegas

Ciudad de México. Año, 2025

CENTRO DE INVESTIGACIÓN EN
CIENCIAS DE INFORMACIÓN GEOESPACIAL, A.C
CentroGeo

Centro Público de Investigación SECIHTI

GENERACIÓN AUTOMATIZADA DE TÓPICOS EN LA PLATAFORMA DE ACCESO A LA INFORMACIÓN PÚBLICA GUBERNAMENTAL (2003-2020)

TESIS

Que para obtener el grado de
Doctor en Ciencias de Información Geoespacial

Presenta

Hermelando Cruz Pérez

Director de Tesis
Dr. Alejandro Molina Villegas

Sinodales
Dr. César Rentería Marín
Dr. Edwyn Aldana Bobadilla
Dra. Silvia Fidelina Fernández Sabido

Julio, 2025

Resumen

El acceso a la información pública es un derecho esencial que permite a la ciudadanía conocer las acciones del Estado y participar en la vida pública. En México, el INAI garantiza este derecho mediante la Plataforma Nacional de Transparencia (PNT). Sin embargo, el crecimiento en las solicitudes ha generado retos para identificar los temas de mayor interés de forma eficiente.

Este estudio propone una metodología automatizada para identificar y categorizar temas en solicitudes de la PNT (2003–2020), empleando técnicas de procesamiento de lenguaje natural y algoritmos genéticos para optimizar parámetros de modelado LDA. Se analizaron más de 2.5 millones de solicitudes, ajustando vocabulario con la Ley de Zipf y parámetros como α, β y K, evaluando la coherencia temática. También se integraron modelos de lenguaje como GPT y LLaMA para mejorar la clasificación automática.

Los resultados revelan que las entidades con más solicitudes fueron Ciudad de México, Estado de México y Jalisco. Los temas sobre servidores públicos tuvieron picos en 2011, 2013 y 2014; mientras que medio ambiente y seguridad mostraron crecimiento sostenido. Temas comerciales y de salud variaron según el contexto.

La metodología desarrollada facilita el análisis temático automatizado, siendo útil para periodistas, investigadores y autoridades. Mejora el acceso y comprensión de la información pública, apoyando decisiones más informadas en las 32 entidades federativas.

Se recomienda aplicar esta metodología a otros periodos y conjuntos de datos. La actualización de modelos es clave para detectar nuevos patrones y fortalecer la transparencia. Ante la desaparición del INAI, estas herramientas cobran mayor relevancia para el monitoreo independiente del derecho de acceso a la información.

Dedicatoria

Dedico esta tesis a mis padres, mi pilar fundamental. Su apoyo incondicional, sus consejos y, sobre todo, su ejemplo de perseverancia y constancia, me han inspirado siempre a luchar por mis metas. Gracias a ellos soy la persona que soy.

A mi familia, por su amor, paciencia y aliento constante en cada etapa de este camino. A mi hija, cuya existencia da sentido y propósito a cada uno de mis esfuerzos; su sonrisa y compañía han sido mi mayor fuente de motivación.

Al CentroGeo, por brindarme las herramientas y el entorno necesarios para desarrollar esta investigación. Agradezco profundamente a las personas que lo integran, por su disposición a colaborar y compartir su conocimiento, alentando la mente de quienes buscan respuestas en los desafíos del presente.

A mi director de tesis, el Dr. Alejandro Molina Villegas, por su experta orientación, paciencia y compromiso académico. Su guía fue esencial para concretar este trabajo, y su ejemplo de rigor y dedicación es una fuente constante de inspiración para seguir formándome y superándome profesionalmente.

Agradezco también el valioso apoyo de la Secretaría de Ciencia, Humanidades, Tecnología e Innovación (Secihti), cuya beca hizo posible la realización de este proyecto de investigación.

Hermelando Cruz Pérez

Índice

Índice de figuras

Índice de Tablas

1. Introducción

1.1 Contexto de las solicitudes de información

El acceso a la información pública es un derecho fundamental que fortalece la transparencia gubernamental y la rendición de cuentas, elementos esenciales en las democracias contemporáneas. A nivel global, la implementación de políticas de acceso a la información ha demostrado ser una herramienta crucial para empoderar a los ciudadanos, mejorar la gobernanza y combatir la corrupción. En México, desde la promulgación de la Ley Federal de Transparencia y Acceso a la Información Pública Gubernamental en 2002, se ha registrado un incremento sostenido en el número de solicitudes de información pública, canalizadas a través de plataformas digitales como SISI, Infomex y, más recientemente, la Plataforma Nacional de Transparencia (PNT).

Entre 2003 y 2020, estas plataformas recibieron más de 2.5 millones de solicitudes, lo que refleja un creciente interés ciudadano en la supervisión de las actividades gubernamentales. Sin embargo, este crecimiento también ha puesto en evidencia una distribución desigual de las solicitudes a lo largo del territorio nacional, lo que indica que las preocupaciones y necesidades de los ciudadanos varían considerablemente según la región. Esta desigualdad en la distribución podría sugerir diferencias en la capacidad de los ciudadanos para ejercer su derecho a la información, lo cual podría tener implicaciones importantes para la equidad y la justicia en la rendición de cuentas gubernamental.

En este contexto, resulta fundamental identificar y comprender las inquietudes específicas de los ciudadanos en diferentes entidades geográficas para desarrollar estrategias de transparencia proactiva y políticas de datos abiertos que se adapten a las necesidades particulares de cada región. Este estudio se enfoca en el análisis de las solicitudes de información presentadas en la PNT durante el periodo 2003-2020, utilizando técnicas avanzadas de procesamiento de lenguaje natural y algoritmos genéticos, con el objetivo de mejorar la accesibilidad, la categorización y el análisis de la información pública.

1.2 Problema de investigación

El problema específico que aborda esta investigación es identificar los temas predominantes que preocupan a los ciudadanos mexicanos en las solicitudes de acceso a la información pública, segmentados por entidad geográfica durante el periodo 2003-2020. Además, busca comprender cómo estas preocupaciones han evolucionado a lo largo del tiempo y cómo varían entre las diferentes regiones del país. Este análisis permitirá mejorar la categorización de los temas y ofrecerá una base sólida para el desarrollo de estrategias y políticas de transparencia adaptadas a las necesidades específicas de cada región. Además, este estudio explorará las posibles desigualdades en el acceso y uso de la información pública, proporcionando una comprensión más profunda de las dinámicas regionales en la demanda de transparencia.

1.3 Hipótesis

A través del análisis automatizado de texto de las solicitudes de acceso a la información pública realizadas por ciudadanos mexicanos entre 2003 y 2020, será posible identificar los temas de interés general de la ciudadanía, así como sus variaciones geográficas y temporales, lo que permitirá apoyar el diseño de estrategias de transparencia más focalizadas, equitativas y alineadas con las necesidades específicas de cada región.

1.4 Objetivo general

Desarrollar una metodología para la identificación y categorización de temas presentes en las solicitudes de acceso a la información pública de la Plataforma Nacional de Transparencia (PNT) durante el periodo 2003–2020, con el fin de mejorar la accesibilidad, análisis y aprovechamiento de la información pública conforme a las necesidades regionales y temáticas de la ciudadanía.

1.5 Objetivos específicos

- Evaluar diferentes enfoques para la identificación automática de tópicos en solicitudes de acceso a la información pública, considerando métricas de coherencia, vocabulario optimizado y modelos de lenguaje para mejorar su interpretación.
- Identificar y clasificar las solicitudes relacionadas con temas ambientales en el marco del Acuerdo de Escazú, con el objetivo de reconocer patrones de interés público y proponer indicadores para su análisis.

- Diseñar una herramienta interactiva que integre análisis temático, visualización estadística y representación geoespacial de solicitudes de información pública, facilitando su consulta por parte de autoridades, investigadores y ciudadanía.
- Comparar el desempeño de la metodología propuesta frente a esquemas tradicionales de clasificación, evaluando su efectividad y aplicabilidad en entornos reales.

1.6 Justificación

La justificación de esta investigación radica en la necesidad de desarrollar una metodología automatizada que utilice técnicas de procesamiento de lenguaje natural y algoritmos genéticos para analizar las solicitudes de información pública presentadas en la PNT durante el periodo 2003-2020. Esta metodología facilitará un acceso más eficiente y permitirá un análisis preciso de la información, ofreciendo además una comprensión más profunda y detallada de las demandas ciudadanas a nivel regional.

Al identificar los temas predominantes por entidad geográfica, esta investigación contribuirá al diseño de estrategias de transparencia proactiva y políticas de datos abiertos más efectivas y alineadas con las necesidades específicas de cada región. Esto permitirá mejorar la rendición de cuentas y promoverá una mayor equidad en el acceso a la información pública, prestando especial atención a las regiones donde la demanda de información es menor pero igualmente significativa. Además, los resultados de este estudio servirán como base para futuras investigaciones en el campo de la transparencia y el acceso a la información pública, y para el desarrollo de recomendaciones orientadas a fortalecer la rendición de cuentas en México.

Esta tesis se compone de seis capítulos. El primero plantea la problemática, las preguntas de investigación, los objetivos y la justificación. El segundo revisa el estado del arte y estudios metodológicos previos. El tercero expone el marco teórico, abordando el acceso a la información, el modelado de temas con LDA y la Ley de Zipf. El cuarto describe la metodología, desde la descarga de datos hasta la visualización de resultados. El quinto presenta los hallazgos, incluyendo la clasificación de tópicos por zonas económicas y la comparación entre modelos como GPT, LLaMA y BETO. El sexto capítulo expone las conclusiones, recomendaciones y futuras líneas de investigación.

2. Marco Teórico

Este capítulo presenta los conceptos necesarios para una comprensión de la automatización de tópicos de las solicitudes de información publica gubernamental en México Se divide en varias secciones clave, incluyendo conceptos teóricos sobre transparencia, acceso a la información, antecedentes del acceso a la información en México, acceso a la información ambiental, la Plataforma Nacional de Transparencia y datos abiertos. Además, se abordan conceptos sobre el modelado de temas, el algoritmo de Latent Dirichlet Allocation (LDA), la ley de Zipf y los algoritmos genéticos.

2.1 Acceso a la información y transparencia

De acuerdo con (McCreadie & Rice, 1999), el derecho al acceso a la información se puede entender como la capacidad de recibir, procesar y utilizar datos de múltiples fuentes, incluyendo documentos, tecnologías de la información y medios de comunicación. Este acceso influye directamente en la calidad de vida y la toma de decisiones, afectado por diversos factores como la tecnología, la economía y la participación ciudadana. Mientras que el acceso a la información pública gubernamental regula la relación entre el Estado, los medios y la sociedad, enfocándose en las normas que delimitan las libertades de expresión e información. Este derecho otorga a las personas la libertad de solicitar y recibir información pública, además de proporcionar mecanismos legales para su protección, como el recurso de revisión ante órganos garantes y la posibilidad de interponer amparos en tribunales federales (López, 2009; Villanueva, 2008).

El derecho de acceso a la información pública es reconocido y respaldado por marcos jurídicos internacionales como la Declaración Universal de Derechos Humanos y el Pacto Internacional de Derechos Civiles y Políticos. Esto permite a cualquier persona compartir, explorar y acceder a la información generada por el gobierno. Sin embargo, este derecho puede estar sujeto a restricciones necesarias para proteger la seguridad nacional, el orden público, la salud o la moral pública, así como los derechos o la reputación de terceros (DUDH, 1948; PIDCP, 2018). En cuanto a la legislación nacional, (Olivos-Fuentes, 2012) destaca que este derecho es reconocido, subrayando la responsabilidad del Estado de garantizar, respetar y promoverlo.

El acceso a la información pública desempeña un doble papel esencial: por un lado, fortalece la autonomía personal y asegura la libertad de expresión y pensamiento; por otro, actúa como un mecanismo importante para el ejercicio de derechos colectivos, permitiendo a los ciudadanos fiscalizar los actos del gobierno y promover la transparencia en la gestión pública (Ramos, 2020). En esta línea, el derecho al acceso a la información se manifiesta en dos dimensiones fundamentales: una activa y otra pasiva. La primera se refiere a la proactividad del Estado en la divulgación de información, mientras que la segunda trata sobre cómo las entidades gubernamentales responden a las solicitudes específicas de información por parte de los ciudadanos, sujetas a restricciones legales justificadas (Fuenmayor E., 2004; Isensee Rimassa & Muñoz Severino, 2010).

En cuanto a la transparencia, definido como la facilidad con la que se accede a información detallada y precisa sobre las actividades gubernamentales, mejorando nuestra habilidad para entender, monitorear y comunicarnos eficazmente (Gutiérrez, 2008). Por su parte, (Hofbauer & Cepeda, 2005, p. 39) agrega que la transparencia implica que las decisiones y costos gubernamentales sean accesibles y claros para el público. Asimismo (Vergara, 2007, p. 17), complementa esta visión al describir la transparencia como un compromiso gubernamental para revelar información a quien la solicite. La transparencia y el acceso a la información pública permiten a los ciudadanos conocer mejor las acciones y decisiones gubernamentales, promoviendo así una gobernanza más abierta y responsable (Perramon, 2013; Sanz Salguero, 2016). En palabras de (Grau, 2006), “La transparencia, en definitiva, puede ayudar a mejorar la calidad democrática de las decisiones y políticas públicas y a potenciar los otros medios de democratización de la administración pública” .

El papel de la tecnología en la transparencia ha permitido la utilización de portales en línea para la divulgación proactiva de datos, Esto mejora la eficiencia administrativa y, además, garantiza un acceso más equitativo a la información. Autores como (Cotino, 2013; Darbishire, 2010) enfatizan cómo Internet ha revolucionado la manera en que se comparte la información gubernamental a través de estos portales, facilitando una participación más amplia y una mejor supervisión por parte de los ciudadanos. De manera análoga, (Cépeda, 2018; Moreno & Castro, 2023) comentan que la divulgación de información de manera anticipada y voluntaria, utilizando medios modernos como Internet, previene la percepción de deshonestidad y mejora la comunicación entre los ciudadanos y el gobierno.

2.1.1 Marco jurídico de acceso a la información

Los antecedentes del acceso a la información pública se remontan a Suecia, donde en 1766 se promulgó la Ley de Libertad de Prensa y el derecho de acceso a los registros públicos. Anders Chydenius, un clérigo sueco-finlandés que también fue diputado y economista, fue el principal impulsor de esta legislación. Esta ley estableció las bases para la apertura de la información gubernamental (Sandoval Ballesteros, 2008).

El concepto moderno de acceso a la información experimentó una transformación significativa tras la proclamación de la Declaración Universal de Derechos Humanos en 1948, que reformuló la libertad de expresión. El artículo 19 de esta declaración establece:

> Todo individuo tiene derecho a la libertad de opinión y de expresión; este derecho incluye el de no ser molestado a causa de sus opiniones, el de investigar y recibir informaciones y opiniones, y el de difundirlas, sin limitación de fronteras, por cualquier medio de expresión(Declaración Universal de Derechos Humanos, 1948, art. 19).

Otro avance en la legislación del derecho al acceso a la información es el Artículo 11 de la (Asamblea Nacional Constituyente de Francia, 1789), que establece:

> La libre comunicación de pensamientos y opiniones es uno de los derechos más valiosos del Hombre; por consiguiente, cualquier ciudadano puede hablar, escribir e imprimir libremente, siempre y cuando responda del abuso de esta libertad en los casos determinados por la Ley.

La libertad de expresión garantiza el acceso a la información pública y la transparencia gubernamental, permitiendo denunciar actos de corrupción y abusos de poder. Este derecho está estrechamente relacionado con el acceso a la información pública, ya que los ciudadanos tienen el derecho de solicitar y obtener información, promoviendo así su participación en la toma de decisiones y la supervisión de los actos gubernamentales.

De igual manera, el artículo 19 del Pacto Internacional de Derechos Civiles y Políticos establece el derecho a la libertad de opinión y de expresión. Este derecho protege a todas las personas de ser molestadas por sus opiniones y garantiza la libertad de buscar, recibir y difundir información e ideas sin restricciones fronterizas, utilizando cualquier medio de su elección (Naciones Unidas, 1966). Sin embargo, el ejercicio de este derecho conlleva responsabilidades y puede estar sujeto a ciertas limitaciones. Estas restricciones deben estar claramente establecidas por la ley y ser necesarias para proteger los derechos y la reputación de los demás, así como para salvaguardar la seguridad nacional, el orden público, o la salud y la moral pública.

Asimismo, la (Organización de los Estados Americanos, 2020) menciona que la Ley Modelo Interamericana 2.0 sobre Acceso a la Información Pública representa un avance importante en el ámbito internacional. Esta herramienta legal establece estándares y buenas prácticas para el acceso a la información pública en América Latina y el Caribe, con el objetivo de fomentar la transparencia, la rendición de cuentas y la participación ciudadana en los asuntos públicos, reconociendo este derecho como fundamental. La ley modelo sirve como base para la creación de leyes nacionales de acceso a la información pública en los países de la región, y propone medidas para garantizar la transparencia en la gestión pública y la publicidad de la información gubernamental, al mismo tiempo que protege los datos personales y la información confidencial.

Con respecto a México, el artículo 6° de la Constitución Política Mexicana establece los principios y bases que rigen el derecho de acceso a la información pública:

> La manifestación de las ideas no será objeto de ninguna inquisición judicial o administrativa, sino en el caso de que ataque a la moral, la vida privada o los derechos de terceros, provoque algún delito, o perturbe el orden público; el derecho de réplica será ejercido en los términos dispuestos por la ley. El derecho a la información será garantizado por el Estado (CPEUM, 1917).

De igual forma, se identifican dos normativas fundamentales referentes a la transparencia: la Ley General y la Ley Federal. La primera aplica a los tres niveles de gobierno y detalla los principios, estructuras y procesos requeridos para la transparencia informativa. La segunda, en cambio, asegura el derecho al acceso a la información mantenida por los Poderes de la Unión y los Órganos Constitucionales Autónomos (Colín & Huesca, 2022). La Ley General de Transparencia se creó mediante la reforma del artículo 6°, la cual fortaleció las atribuciones del INAI (Instituto Nacional de Transparencia, Acceso a la Información y Protección de Datos Personales) y estableció estándares de transparencia y acceso a la información en todo el país. Además, la ley obliga al Congreso de la Unión a crear una Ley General de Transparencia que se aplique en todos los niveles de gobierno (Bautista-Farías, 2015).

La Ley General de Transparencia y Acceso a la Información Pública (DOF, 2015) establece los principios y procedimientos para garantizar el acceso a la información en posesión de diversas entidades y organismos del gobierno, así como de personas físicas, morales y sindicatos que reciben recursos públicos o realizan actos de autoridad. Esta ley, de orden público y aplicable en todo el territorio nacional, regula el derecho de acceso a la información en concordancia con el artículo 6º de la Constitución Política de los Estados Unidos Mexicanos. Todo este marco jurídico mencionado otorga a los ciudadanos, residentes o partes interesadas el derecho a acceder a la información bajo el resguardo del gobierno sin necesidad de demostrar interés legal (Ackerman & Sandoval, 2008).

2.1.2 Problemas de implementación y cumplimiento

Los ciudadanos en México enfrentan cuatro desafíos principales al solicitar información pública. El primero es el desconocimiento de la Ley de Acceso a la Información Pública y el procedimiento para solicitarla. El segundo es la limitada competencia en el uso de tecnologías de la información y la comunicación. El tercero es la falta de usabilidad de la PNT, debido al desconocimiento de su funcionamiento y a la falta de accesibilidad. El cuarto desafío es el desconocimiento del derecho al acceso a la información y la identificación de los sujetos obligados (Aguilar-Arévalo & Ramírez-Montaño, 2019).

Otras problemáticas identificadas por (Cueto, 2017) incluyen la existencia de diferentes procedimientos para sancionar incumplimientos, la falta de evaluación y capacitación adecuada, la ausencia de leyes específicas de protección de datos en muchas entidades federativas, la asignación desigual de presupuesto a los órganos garantes, la concentración de solicitudes de información en pocas entidades y un bajo porcentaje de solicitudes con recursos de revisión. Además, persisten problemas como la opacidad de los actos públicos, la prevalencia de la corrupción y la ilegalidad (Pérez & Cecilia, 2019). Otro desafío significativo que enfrentan las personas al ejercer el derecho al acceso a la información pública es la necesidad de competencias digitales adecuadas (Ruiz, 2009), así como la falta de concientización y el manejo indebido de datos sensibles (Osollo, 2021).

Para superar estos obstáculos, se requiere un mayor conocimiento sobre la ley y la plataforma, así como un mejor aprovechamiento de la tecnología para acceder a la información pública. Tal como señala (Fierro, 2021), para mejorar el ejercicio del acceso a la información pública gubernamental, es necesario definir los objetivos y determinar qué información de cada sujeto obligado es de interés público y en qué formatos presentarla. Además, para que este derecho sea efectivo, es necesario superar la brecha digital. Es esencial que, desde el momento en que se genera la información, esta sea publicada en los portales correspondientes para asegurar el acceso público y oportuno (Martínez & de Mingo, 2018).

2.1.3 Impacto y relevancia social

El acceso a la información pública es esencial para garantizar la transparencia y la rendición de cuentas en la toma de decisiones gubernamentales. Sin este acceso, ciudadanos e investigadores no pueden comprender plenamente qué decisiones se toman, quiénes están involucrados y qué información se utiliza (Héritier, 2003). Esto impide a los periodistas e investigadores profundizar más allá del discurso oficial y evaluar el impacto de las decisiones en la sociedad (Walby & Larsen, 2012). Además, (Yannoukakou & Araka, 2014) enfatizan que el acceso a la información es fundamental para la transparencia, la rendición de cuentas y la gobernanza participativa, además de fomentar la innovación y el crecimiento económico mediante datos abiertos.

Este derecho también facilita el ejercicio de otros derechos, como la participación ciudadana y la libertad de expresión, subrayados por la Declaración Universal de los Derechos Humanos (Valencia, 2011). De este modo, el acceso a la información pública empodera a los ciudadanos y promueve una participación activa y consciente en la gestión pública y los procesos democráticos (Curich, 2016).

2.1.4 Acceso a la información ambiental

El 4 de marzo de 2018, América Latina y el Caribe aprobaron el Acuerdo de Escazú, fundamentado en el Principio 10 de la Declaración de Río de 1992. Este acuerdo, resultado de dos años de preparación y nueve reuniones de negociación lideradas por Chile y Costa Rica, asegura los derechos de acceso a la información, participación pública y acceso a la justicia en asuntos ambientales (CEPAL, 2018). México formalizó su compromiso con este tratado al depositar el instrumento de ratificación el 22 de enero de 2021, y el acuerdo entró en vigencia el 22 de abril del mismo año (Muñoz, 2023).

El Acuerdo de Escazú proporciona un marco internacional para avanzar en la transparencia y la participación pública en asuntos ambientales en México. Sin embargo, hemos sido testigos de un alarmante aumento de la violencia y los asesinatos. Según (Witness, 2021), durante el año 2020 se incrementaron los ataques dirigidos a personas que defienden la tierra y el medio ambiente, así como a comunidades indígenas. La explotación forestal ha sido responsable de aproximadamente un tercio de estos ataques en todo el país.

En consonancia con lo anterior, el Centro Mexicano de Derecho Ambiental (CEMDA) reporta que el año 2021 se destacó tristemente como uno de los más violentos en la defensa del patrimonio natural. Las agresiones más frecuentes incluyen intimidación, hostigamiento, amenazas, agresiones físicas, homicidio y desaparición. En cuanto a la ubicación geográfica de estos ataques, Oaxaca ocupa el lamentable primer lugar en la lista de estados con mayor número de ataques letales en el país. Guerrero le sigue de cerca como la segunda entidad con más agresiones letales, compartiendo esta preocupante estadística con Sonora, donde al menos cuatro personas defensoras fueron víctimas de asesinato (González et al., 2022).

El Acuerdo de Escazú se convierte en una herramienta esencial para enfrentar la creciente ola de violencia y proteger a quienes trabajan incansablemente para cuidar nuestro entorno natural. Además, ofrece un marco para fomentar la transparencia y la participación pública en temas ambientales, lo cual es fundamental para avanzar hacia un futuro más sostenible.

2.1.5 Plataforma Nacional de Transparencia

La Plataforma Nacional de Transparencia es una herramienta fundamental para ejercer los derechos de acceso a la información y la protección de datos personales en posesión de los sujetos obligados a través de medios electrónicos, garantizando uniformidad y sirviendo como repositorio de información obligatoria a nivel nacional (DOF, 2016). La Plataforma Nacional, accesible en [1], está compuesta por varios sistemas que desempeñan funciones específicas. Según (Ford, 2016), la PNT cuenta con cuatro subsistemas principales.

[1] https://www.plataformadetransparencia.org.mx/

- Sistema de solicitudes de acceso a la información: Este sistema permite la presentación de solicitudes de acceso a la información pública y de acceso, rectificación, cancelación y oposición de datos personales (derechos ARCO). Los solicitantes pueden realizar su solicitud a través de la plataforma, conocer el estatus de la misma, dar seguimiento al proceso de respuesta y recibir la respuesta en el medio que prefieran. Los servidores públicos pueden asignar la solicitud a la unidad de transparencia del sujeto obligado, la cual podrá ser turnada al área que posea la información para preparar la respuesta correspondiente.
- Sistema de gestión de medios de impugnación: Este sistema permite a los solicitantes inconformes con la respuesta del sujeto obligado interponer un recurso de revisión ante el organismo garante de transparencia de la entidad federativa o ante el INAI para el caso de instituciones federales. La plataforma facilita la interposición de este tipo de recursos y el organismo garante puede acceder al registro de la atención de la solicitud y a la documentación generada en el proceso de respuesta.
- Sistema de comunicación entre organismos garantes y sujetos obligados: Este sistema permite la comunicación entre organismos garantes y sujetos obligados en relación con el proceso de atención de solicitudes de acceso a la información. Permite el trámite del recurso de revisión y la interposición de un recurso de inconformidad ante el INAI.
- Sistema de gestión de obligaciones de transparencia: Este sistema permite la gestión de las obligaciones de transparencia de los sujetos obligados, es decir, la publicación de información de oficio. La plataforma facilita la actualización y la consulta de la información por parte de la ciudadanía.

2.1.6 Datos abiertos de la Plataforma Nacional de Transparencia

Los datos abiertos ofrecen beneficios como la toma de decisiones basada en evidencia, intervenciones más efectivas y una asignación transparente de recursos públicos. Para lograr estos resultados, es necesario que los datos se generen y publiquen de manera oportuna, coincidiendo con el ciclo de diseño de políticas. Esto implica la captura y publicación automatizada de información en los sitios web de las entidades responsables, así como la publicación en tiempo real en un portal central. De esta manera, los datos estarán disponibles de manera segura y catalogados desde su fuente, permitiendo su uso efectivo en la toma de decisiones (ONU, 2014). Conforme a la Ley General de Transparencia y Acceso a la Información Pública, (DOF, 2015), los datos abiertos son datos públicos accesibles en línea, reutilizables y redistribuibles.

Estos datos deben cumplir con características como: accesibilidad para una amplia audiencia, integridad y descripción detallada, disponibilidad gratuita, no discriminación en el acceso, actualización constante, conservación de versiones anteriores, provisión directa desde la fuente, legibilidad por máquinas, citación de la fuente de origen y disponibilidad en formatos abiertos como CSV, XML, JSON y TXT. La PNT cuenta con una sección de datos abiertos que incluye las solicitudes de información realizadas a las dependencias de la administración pública federal, junto con sus correspondientes respuestas en formatos de datos abiertos (SQL, XML, CSV y JSON). La información puede descargarse por periodos años, o por sectores específicos, como Comunicaciones y Transportes, Aportaciones a la Seguridad Social, Defensa Nacional, entre otros.

2.2 Modelado de Tópicos

Según (Jelodar et al., 2019; Li & Yamanishi, 2003), los modelos de tópicos son herramientas importantes en la minería de texto y el procesamiento de lenguaje natural, que ayudan a identificar los temas principales en un conjunto de documentos. Funcionan seleccionando palabras de posibles grupos temáticos repetidamente hasta encontrar la distribución más probable. Aunque no interpretan el significado exacto de las palabras, ofrecen una comprensión valiosa de grandes colecciones de textos. Estos modelos son clave para crear resúmenes automáticos, identificar patrones y tendencias en los datos, y entender mejor el contenido y la estructura temática de la información, incluyendo texto, audio y vídeo.

En concordancia con (Mohr & Bogdanov, 2013), el modelado de temas es una técnica de análisis de texto que automatiza la identificación de estructuras ocultas en grandes colecciones de documentos, útil en disciplinas como las ciencias sociales y las humanidades. A diferencia de los métodos tradicionales que requieren una codificación manual exhaustiva, este enfoque utiliza algoritmos que, con mínima intervención humana, determinan y distribuyen temas basándose en la probabilidad de aparición de las palabras. Aunque no es infalible, cuando se aplica de manera cuidadosa, el modelado de temas puede ofrecer interpretaciones plausibles y coherentes de los textos.

2.2.1 Modelado probabilístico de temas

$$P(w_i) = \sum_{j=1}^{T} P(w_i|z_i = j) \cdot P(z_i = j) \tag{2.1}$$

La ecuación 2.1 representa la **probabilidad marginal** de observar una palabra w_i en el corpus, considerando su posible asociación con distintos tópicos. Esta probabilidad se calcula como la suma ponderada de dos componentes clave en el modelado de temas:

- $P(w_i|z_i = j)$: la probabilidad de que la palabra w_i aparezca dado el tópico j. Esta es la distribución de palabras dentro de cada tópico.
- $P(z_i = j)$: la probabilidad de que el tópico j sea seleccionado, es decir, la distribución de tópicos.

La suma $\sum_{j=1}^{T}$ recorre todos los posibles tópicos j (de 1 a T), y permite obtener la probabilidad total de que una palabra aparezca, considerando todas las temáticas del modelo.

Esta formulación es fundamental en modelos generativos de tópicos como LDA, ya que permite descomponer un corpus complejo en estructuras latentes de temas, facilitando el análisis temático de grandes volúmenes de texto.

2.2.2 Técnicas de clasificación de tópicos

El estudio de (Abdelrazek et al., 2023) clasifica el modelado de tópicos en cuatro categorías: algebraicos, difusos, bayesianos probabilísticos y neuronales. Los algebraicos aproximan matrices término-documento, mientras que los difusos emplean grados de certeza para asignar palabras a grupos. Los bayesianos implementan procesos generativos mediante gráficos acíclicos dirigidos y ajustan datos mediante inferencia. Los modelos neuronales, por otro lado, recurren a la optimización para abordar inferencias complejas, esta clasificación se detalla en la Tabla 2.1.

Tabla 2.1. Clasificación de modelos de tópicos.

Categoría	Descripción	Ventajas	Desventajas	Ejemplos
Algebraico	Descomponer la matriz término-documento, luego encontrar una aproximación de bajo rango.	Simple, intuitivo y computacionalmente relativamente eficiente. Puede manejar documentos cortos.	No proporciona una base estadística sólida. No define un modelo generativo de datos.	LSA, NMF
Difuso	Extraer tópicos de los documentos usando lógica difusa.	Puede manejar dispersión en textos cortos (ej. tweets).	Mayor enfoque en datos médicos.	Modelos de conjuntos difusos
Bayesiano probabilístico	Define un proceso generativo a través de un modelo gráfico bayesiano.	Intuitivo, extensible e interpretable.	Inferencia compleja con modelos grandes.	LDA, Model Word Order
Modelos neuronales	Sustituye la inferencia posterior con optimización.	Flexible, coherente, escalable.	Menor interpretabilidad.	BERTopic, ETM

Fuente: Adaptado de Abdelrazek et al. (2023).

Modelos algebraicos

Los modelos algebraicos en el modelado de tópicos usan técnicas matemáticas para descomponer matrices grandes y así simplificar y descubrir las estructuras ocultas en los datos. A continuación, se presentan los más comunes.

- Latent Semantic Analysis (LSA)

  LSA puede entenderse como un modelo que describe cómo adquirimos y usamos el conocimiento a través de procesos computacionales y representaciones subyacentes. Funciona aplicando cálculos estadísticos en un gran conjunto de textos para extraer y representar el significado contextual de las palabras (Landauer et al., 1998). Según (Steinberger & Jezek, 2004), usar Latent Semantic Analysis (LSA) para la detección de temas en textos presenta algunos desafíos. Es necesario ajustar el número de oraciones en el resumen, y un alto número de dimensiones puede reducir la relevancia de los temas. Además, la matriz de términos por oraciones puede ser muy grande, lo que requiere muchos recursos computacionales para la descomposición en valores singulares. Por último, al no considerar la estructura del texto, LSA puede generar resultados inexactos o incompletos en la detección de temas.

- Non-negative matrix factorization

  Según (Lee & Seung, 1999), el uso de NMF (Factorización Matricial No Negativa) ha permitido analizar el contenido semántico de los documentos modelando la aparición de palabras a través de características semánticas ocultas. Estas características identifican patrones de coexistencia de palabras, facilitando el descubrimiento de temas. La investigación de (Carbonetto et al., 2021) destaca cómo los algoritmos(NMF) mejoran la estimación de parámetros en modelos de temas. NMF simplifica el proceso de optimización al no requerir las restricciones de suma a uno típicas de estos modelos, permitiendo cálculos más eficientes.

Modelos Difuso

La lógica difusa es un enfoque matemático que maneja la incertidumbre y la imprecisión en el razonamiento y la toma de decisiones, reconociendo que las verdades no siempre son absolutas, sino que pueden ser parcialmente ciertas o falsas. Esta capacidad para asignar grados de verdad permite su aplicación en varios campos como inteligencia artificial, sistemas de control, decisiones estratégicas y modelado de tópicos (Lai & Chen, 2023). De manera similar, (Rashid et al., 2019) introduce un enfoque para el modelado de tópicos desde una perspectiva difusa, para solucionar problemas de ruido y falta de datos en textos cortos. Este método, denominado Modelado de Tópicos Difuso (FTM), utiliza el modelo de Bolsa de Palabras (BOW) para generar frecuencias de términos locales y globales, mejorando la precisión y coherencia de los tópicos extraídos.

Modelos Bayesiano probabilístico

- Asignación Latente de Dirichlet

  La Asignación Latente de Dirichlet (LDA) es un algoritmo no supervisado de análisis de tópicos que se utiliza para identificar los temas subyacentes en un conjunto de documentos. Este algoritmo asume que cada documento está compuesto por una combinación de temas latentes y que cada tema está formado por un conjunto de palabras (Blei et al., 2003).

  LDA busca representar los documentos en términos de una distribución de probabilidad sobre un conjunto finito de temas latentes, donde cada tema está representado por una distribución de probabilidad sobre el vocabulario. Para generar un nuevo documento, LDA primero selecciona una mezcla de temas de acuerdo con la distribución de temas del documento. Luego, para cada palabra en el documento, se selecciona un tema según la mezcla elegida y se elige una palabra de ese tema de acuerdo con su distribución de probabilidad en el vocabulario.

Según (Rieger et al., 2020, p. 2), en un modelo de Latent Dirichlet Allocation (LDA), los valores de α , β y el número de temas son parámetros clave que controlan el comportamiento del algoritmo y afectan la forma en que se generan los temas. A continuación se describen cada uno de estos parámetros:

α: Controla la distribución de temas en los documentos. Un valor alto de α indica que los temas están distribuidos de forma uniforme a lo largo de los documentos, mientras que un valor bajo de α indica que algunos temas están más concentrados en ciertos documentos.

β: Controla la distribución de palabras en los temas. Un valor alto de β sugiere que las palabras están distribuidas uniformemente a lo largo de los temas, mientras que un valor bajo de β indica que algunas palabras están más concentradas en ciertos temas.

El parámetro K especifica cuántos temas se generarán en el modelo y afecta directamente la interpretabilidad de los temas obtenidos. A medida que K aumenta, el modelo se vuelve más complejo y los temas generados resultan más difíciles de interpretar.

Según (Calistus et al., 2024), en el modelado de temas, la Asignación Latente de Dirichlet (LDA) enfrenta desafíos importantes, como determinar el número óptimo de temas para evitar generalización excesiva o temas confusos. La eficacia de LDA depende de la correcta configuración de los parámetros α y β, que regulan la distribución de temas y palabras; una mala configuración puede comprometer la estructura latente de los datos. Además, la interpretabilidad y la adaptación temporal del modelo son retos adicionales, especialmente en análisis complejos donde es importante entender la evolución de los temas a lo largo del tiempo.

El trabajo realizado por (Blei et al., 2003) introdujo el concepto de Asignación de Dirichlet Latente (LDA), utilizando la distribución de Dirichlet. En la ecuación (2.2), los parámetros de esta distribución controlan la mezcla de temas dentro de los documentos. Aquí, Γ es la función Gamma, que extiende la función factorial a los números reales, y K representa el número total de temas en el corpus, mientras que α es un hiperparámetro que influye en la forma de la distribución, afectando la concentración y diversidad de la mezcla de temas.

La distribución de Dirichlet, definida por la ecuación (2.2), es clave en el proceso generativo del LDA, ya que especifica las probabilidades a priori de las distribuciones de temas en los documentos. El término p_j representa la probabilidad asignada a cada tema j, con el exponente $\alpha - 1$ modulando la influencia de cada tema según el hiperparámetro α. El objetivo es determinar los valores de probabilidad de la distribución para que los documentos originales puedan ser generados.

$$Dir(\alpha) = \frac{\Gamma(\alpha K)}{\prod_{j=1}^{K} \Gamma(\alpha)} \prod_{j=1}^{K} p_j^{\alpha-1} \tag{2.2}$$

En la ecuación para la distribución de Dirichlet dentro del LDA:

- $Dir(\alpha)$: Distribución de Dirichlet parametrizada por α.
- α: Controla la uniformidad de los temas; valores más altos significan una distribución más uniforme.
- K: Número de temas.
- Γ: Función Gamma, utilizada para normalización.
- p_j: Probabilidad del tema j.
- $\prod$: Producto de las probabilidades de los temas, influenciado por α.

■ Correlated Topic Model

(Blei & Lafferty, 2007) introdujeron el modelo de Tópicos Correlacionados (CTM), un modelo jerárquico diseñado para colecciones de documentos que utiliza una mezcla de componentes compartidos entre documentos y proporciones específicas para cada uno. Este enfoque permite representar múltiples tópicos dentro de un mismo documento, capturando así la heterogeneidad y los patrones latentes en los datos. A diferencia del modelo de Asignación LDA, el CTM emplea la distribución normal logística, lo que introduce una estructura de covarianza entre tópicos y mejora el rendimiento predictivo del modelo. Esto permite detectar relaciones complejas entre tópicos de manera más efectiva.

Modelos de temas neuronales

Estos modelos se dividen en dos subcategorías principales: modelos de incrustación neuronal y modelos variacionales. A continuación, se presenta una descripción de cada uno.

Neural Embedding:

■ LDA2VEC: De acuerdo a (Moody, 2016), este modelo fusiona el modelado de tópicos con incrustaciones de palabras para generar representaciones documentales no supervisadas y coherentes. Entrena vectores de palabras, tópicos y documentos en un espacio común que mantiene regularidades semánticas y ofrece asignaciones de documentos a tópicos claras y compactas.

- Bertopic: Es modelo de última generación para el modelado de temas que combina la técnica de BERT con el modelado de temas tradicional. BERTopic crea incrustaciones de documentos utilizando modelos de lenguaje basados en transformadores preentrenados, agrupa estas incrustaciones y, posteriormente, emplea un procedimiento TF-IDF basado en clases para formar las representaciones de los tópicos (Grootendorst, 2022).

Varational:

- TopicRNN : De acuerdo con (Dieng et al., 2016), el modelo TopicRNN fusiona redes neuronales recurrentes (RNN) con tópicos latentes para capturar tanto la estructura semántica local como global de los documentos. Mientras las RNN abordan las relaciones sintácticas y semánticas inmediatas, los tópicos latentes aportan un entendimiento más amplio y descontextualizado. TopicRNN, que se aprende de manera integral, mejora la predicción de palabras y funciona eficazmente como extractor de características no supervisado, ofreciendo una alternativa robusta a métodos tradicionales como la asignación de Dirichlet latente.
- Autoencoder Variacional (VAE): Un VAE combina un modelo generativo de variables latentes con un modelo de inferencia para estimar la distribución de estas variables. Optimizado mediante el ascenso de gradiente estocástico en el límite inferior de la evidencia (ELBO), el VAE es una herramienta esencial en el modelado probabilístico y la creatividad artificial, con implementaciones disponibles en las principales bibliotecas de aprendizaje profundo. En el análisis de tópicos, los VAE descubren estructuras temáticas latentes en textos, facilitando la clasificación y el análisis de grandes colecciones documentales(Kingma, Welling et al., 2019).

2.2.3 Evaluación de tópicos

La evaluación de tópicos es importante para determinar si un modelo ha identificado temas relevantes en un corpus. Uno de los métodos más comunes es medir la coherencia, que evalúa qué tan bien están agrupadas las palabras dentro de un tema. Se comienza seleccionando palabras clave basadas en su frecuencia y relevancia, luego se analiza cómo estas palabras se relacionan entre sí, considerando su similitud semántica. Finalmente, se utiliza un cálculo estadístico para asignar una puntuación de coherencia, lo que permite evaluar la efectividad del modelo en identificar temas coherentes (Röder et al., 2015).

Como se detalla en (Newman et al., 2010), el papel de la coherencia en la modelización de temas es instrumental para determinar la interpretabilidad de los temas por analistas humanos. Caracterizando temas a través de las palabras más probables, la puntuación de coherencia evalúa el grado de similitud entre estas palabras, sirviendo como un indicador crucial de coherencia temática. Una de las métricas más utilizadas es la puntuación de coherencia de UMass, Ecuación 2.3. Esta métrica calcula la frecuencia con la que dos palabras aparecen juntas en el corpus.

$$C_{UMass} = \sum_{i=2}^{N} \sum_{j=1}^{i-1} \log \frac{D(w_i, w_j) + \epsilon}{D(w_j)} \tag{2.3}$$

Los componentes de la ecuación de coherencia UMass se delinean a continuación, explicando sus roles en la evaluación de la coherencia temática dentro del ámbito de la modelización de temas:

- C_{UMass}: La puntuación final de coherencia UMass evalúa la coherencia semántica entre las palabras del tema.
- N: El conteo de palabras únicas en el tema evaluado.
- w_i, w_j: Las palabras indexadas del tema para comparación por pares.
- $D(w_i, w_j)$: Frecuencia documental de ambos w_i y w_j, indicando co-ocurrencia.
- ϵ: Un valor positivo insignificante para evitar logaritmos indefinidos.
- $D(w_j)$: Frecuencia documental de w_j, el conteo de ocurrencias de w_j.
- $\log$: Función logarítmica, normalizando la puntuación de coherencia.
- $\sum$: Suma sobre pares de palabras.

Crucialmente, C_{UMass} cuantifica la cohesión temática sumando sobre pares de palabras para evaluar su co-ocurrencia ($D(w_i, w_j)$) y frecuencias individuales ($D(w_j)$), ajustadas por ϵ para asegurar estabilidad matemática. Este cálculo, respaldado por la relación logarítmica, ofrece insights sobre las conexiones semánticas de las palabras. La puntuación de coherencia agregada, derivada del promedio de estas puntuaciones por pares, refleja la armonía temática anclada en las frecuencias del corpus que originalmente informaron los modelos de temas, este enfoque asegura un análisis de temas matizado y sensible al contexto (Stevens et al., 2012).

2.2.4 La ley de Zipf en la optimización del vocabulario

La distribución de frecuencia de palabras en los idiomas humanos sigue un fenómeno fundamental en los estudios lingüísticos, conocido como la Ley de Zipf presentada en la Fórmula (2.4).

Esta ley sigue una regularidad sistemática y universal: ciertas palabras, como “de”, “la”, y “que”, ocurren con alta frecuencia y dominan las palabras funcionales, mientras que los términos semánticamente específicos del documento aparecen de manera esporádica (Piantadosi, 2014). Este patrón muestra una consistencia notable a través de diversos idiomas naturales y contextos, influyendo significativamente en el procesamiento del lenguaje natural. Estas propiedades se utilizan para automatizar la distinción entre palabras funcionales y términos relevantes para el modelado de temas.

La Ley de Zipf articula esta distribución como:

$$f(r) \propto \frac{1}{r^{\alpha}} \tag{2.4}$$

donde:

- $f(r)$ representa la frecuencia del término con rango r.
- α es un exponente, que en los idiomas naturales típicamente está cerca de 1.
- $\propto$ denota proporcionalidad, indicando que la frecuencia de una palabra es inversamente proporcional a su rango en la lista de frecuencias, elevada a la potencia de α.

2.2.5 Algoritmo genético

Los algoritmos genéticos son técnicas de optimización basadas en los principios de selección natural y genética. Estos métodos combinan la supervivencia de las mejores soluciones con la introducción de nuevas variaciones. En cada generación, se crea una nueva población de individuos a partir de segmentos de los individuos más destacados de la generación anterior. Periódicamente, se añaden variaciones para mejorar la adaptación y explorar nuevas soluciones (Goldberg, 1989).

En la década de 1960, John Holland fue pionero en los algoritmos genéticos (GA) y trabajó con sus estudiantes y colegas en la Universidad de Michigan durante las décadas de 1960 y 1970. A diferencia de otros enfoques como las estrategias de evolución y la programación evolutiva, el objetivo principal de Holland no era diseñar algoritmos específicos para resolver problemas particulares. Su enfoque se centraba en investigar formalmente el proceso de adaptación en la naturaleza y desarrollar métodos para incorporar estos mecanismos en sistemas informáticos.

El trabajo de Holland y sus aportaciones fundamentales son expuestos en detalle en el libro de Melanie Mitchell, *Una introducción a los algoritmos genéticos*, donde se resalta cómo estos algoritmos pueden optimizar funciones a pesar de su carácter aleatorio. En esta misma obra, se subraya que los algoritmos genéticos son especialmente robustos por cuatro razones: (1) emplean una representación codificada de los parámetros, (2) operan sobre una población de soluciones en lugar de un único punto, (3) dependen únicamente de la función objetivo sin requerir derivadas, y (4) dirigen la búsqueda mediante reglas probabilísticas en lugar de deterministas (Mitchell, 1998).

Según Michalewicz (Zbigniew, 1996), un algoritmo genético consta de varios componentes esenciales: una representación genética para codificar las soluciones, un procedimiento para generar la población inicial, una función de evaluación para calificar las soluciones, operadores genéticos que alteran la composición de las generaciones futuras, y ajustes de parámetros como el tamaño de la población y las probabilidades de aplicación de los operadores genéticos. La mayoría de los algoritmos genéticos comparten componentes fundamentales como poblaciones de cromosomas, selección basada en la aptitud, procesos de cruce y mutaciones aleatorias (Goldberg, 1989).

2.2.6 Antecedentes de los transformers en PLN

Uno de los antecedentes importantes es el trabajo de Morin y Bengio (Morin & Bengio, 2005), que propone acelerar el entrenamiento y reconocimiento de modelos lingüísticos basados en redes neuronales. Utiliza una descomposición jerárquica basada en la taxonomía de WordNet, que organiza conceptos en un grafo similar a un árbol. El objetivo principal es crear una versión más rápida del modelo neural probabilístico del lenguaje, reduciendo el tiempo de entrenamiento en función del tamaño del vocabulario. Este enfoque mejora significativamente la velocidad de entrenamiento y reconocimiento, capturando regularidades semánticas y sintácticas. Sus aplicaciones incluyen el modelado del lenguaje, reconocimiento del habla y traducción automática.

El trabajo de Mikolov et al. (Mikolov et al., 2013) introduce el modelo skip-gram continuo, un método eficiente para aprender representaciones vectoriales de palabras a partir de texto no estructurado. Este modelo no requiere multiplicaciones matriciales densas, lo que lo hace muy eficiente. Utiliza submuestreo de palabras frecuentes para mejorar la calidad y velocidad del entrenamiento, y el algoritmo de muestreo negativo para obtener representaciones vectoriales precisas. Este enfoque mejora significativamente la eficiencia del entrenamiento y la calidad de las representaciones, con aplicaciones en tareas como el razonamiento analógico y la bolsa continua de palabras.

Un avance destacado en el procesamiento del lenguaje natural es el modelo Transformer (Vaswani et al., 2017). Esta arquitectura de red neuronal se basa en mecanismos de autoatención para capturar dependencias globales entre la entrada y la salida, eliminando la necesidad de capas recurrentes o convoluciones. Esto permite una mayor paralelización y tiempos de entrenamiento más rápidos. El mecanismo de atención del Transformer mejora el rendimiento y la calidad de la traducción, y también es eficaz en otras tareas como el análisis sintáctico y el reconocimiento de voz. Su versatilidad se extiende a diferentes modalidades, incluyendo imágenes, audio y video, ampliando su aplicabilidad en el procesamiento del lenguaje natural.

2.2.7 Embeddings

Los embeddings o word embeddings son representaciones numéricas densas de palabras en un espacio de baja dimensionalidad. Estas representaciones se obtienen entrenando modelos de redes neuronales con grandes conjuntos de texto. Los word embeddings están diseñados para capturar similitudes entre palabras: palabras que aparecen en contextos similares se agrupan cerca unas de otras en el espacio proyectado. Estas representaciones son efectivas como características en diversas tareas de procesamiento del lenguaje natural, también se les conoce como embeddings neurales o simplemente embeddings (Levy & Goldberg, 2014, p. 171).

2.2.8 BETO

De acuerdo con (Cañete & et al., 2020, p. 1), BETO es un modelo basado en BERT, entrenado desde cero con un corpus extenso en español de tamaño comparable al utilizado para BERT-Base. Comparte los mismos principios de diseño y emplea la técnica de Máscara de Palabras Completa (Whole Word Masking), lo que lo convierte en una base sólida para evaluar el desempeño de modelos en tareas de procesamiento del lenguaje en español.

3. Estado del Arte

A continuación, se presentan algunos estudios previos que se centran en el acceso a la información, el uso de LDA y la optimización de sus parámetros.

3.1 Acceso a la información

- La investigación realizada por (Coria & López, 2024) propone una guía metodológica para implementar técnicas de minería de datos en el contexto de la Plataforma Nacional de Transparencia, enfocándose en particular en el Sistema de Portales de Obligaciones de Transparencia. La relevancia de este trabajo se destaca por la creciente necesidad de procesar y analizar datos públicos de manera confiable, crucial para una mejor comprensión de dinámicas sociales complejas.
- En los estudios de (Bagozzi et al., 2016, 2019), los autores emplearon el modelo de Asignación de Dirichlet Latente Supervisada (sLDA), una extensión del modelo LDA que incorpora variables de salida supervisadas. Este enfoque fue utilizado para analizar la capacidad de respuesta del gobierno mexicano a las solicitudes de información pública entre 2003 y 2015, permitiendo identificar los temas más asociados con las respuestas o la falta de ellas por parte del gobierno.
- De acuerdo con los hallazgos de (Berliner et al., 2018) el estudio de más de un millón de solicitudes de información pública gubernamental de México, utilizando métodos no supervisados para categorizarlas basándose en su diversidad temática. Este análisis muestra la diversidad de temas de las solicitudes, que abarcan desde cuestiones gubernamentales, personales y políticas, destacando las preocupaciones de los ciudadanos por temas ambientales y de violencia.
- En el trabajo de (Berliner et al., 2020) se analizan las respuestas del gobierno a las solicitudes de información realizada por los ciudadanos en la plataforma INFOMEX. Los autores sostienen que si una solicitud de información proviene de una región que apoya al partido en el poder, es más probable que reciba una respuesta favorable y rápida respecto a otras solicitudes.

3.2 Optimización de hiperparametros (LDA)

La implementación del Algoritmo de Gibbs Condicionado (CGS) para optimizar el modelo de LDA ha demostrado mejorar la eficacia en la clasificación de temas. Sin embargo, alcanzar una precisión óptima requiere múltiples iteraciones y una selección cuidadosa de los parámetros α y β, ya que una elección incorrecta de estos puede afectar tanto la precisión del modelo como la determinación del número óptimo de temas (Subeno, 2017). A continuación, se presentan trabajos que se han realizado para optimizar los hiperparámetros del modelo LDA.

- El trabajo realizado por (Yarnguy & Kanarkard, 2018) Llevaron a cabo la parametrización automática de los hiperparámetros α y β utilizando un enfoque de optimización de colonia de hormigas (ACO) para determinar los valores óptimos. Para la evaluación, utilizaron datasets estándar de UCI (KOS, NIPS, ENRON) para estimar el modelo de temas. Los resultados de su investigación indican que el modelo LDA con parámetros ajustados mediante ACO tiene un mejor rendimiento al ser evaluado por la puntuación de perplejidad.
- Por su parte, (Pathik & Shukla, 2020) utilizaron un algoritmo de recocido simulado para optimizar los hiperparámetros del modelo LDA. Realizaron una evaluación empírica utilizando conjuntos de datos de revisiones de clientes. Los resultados de la experimentación muestran que SA-LDA ofrece un mejor rendimiento cuando se evalúa con la puntuación de coherencia.
- El estudio de (Subeno, 2017) explora cómo determinar el número óptimo de temas en un corpus utilizando la técnica de máxima verosimilitud y el Método de la Longitud Mínima de Descripción (MDL). Realizaron experimentos con artículos de noticias y descubrieron que los parámetros α y β juegan un papel crucial en la determinación del número óptimo de temas. Además, encontraron que el volumen de documentos no afecta los tiempos de computación, pero el número de palabras sí tiene un impacto considerable.
- En el trabajo de (Hasan et al., 2021) introducen dos metodologías, Coherencia Absoluta Normalizada (NAC) y Perplejidad Absoluta Normalizada (NAP), diseñadas para optimizar la determinación del número óptimo de temas. El estudio cuestiona la eficacia de los enfoques convencionales y demuestra que establecer un número excesivo de temas puede ser contraproducente, ya que aumenta el tiempo de procesamiento sin mejorar la precisión. Los resultados sugieren que, a pesar de las limitaciones reconocidas que serán abordadas en investigaciones futuras, las metodologías propuestas mejoran significativamente la robustez en la extracción de características de textos extensos.

- En el análisis realizado por (Zhao et al., 2015) proponen un enfoque heurístico basado en la tasa de cambio de la perplejidad (RPC) que podría simplificar el desarrollo de modelos de temas, permitiendo un ajuste más preciso y eficiente. Los resultados muestran que el método basado en RPC es estable, preciso y efectivo, validado a través de experimentos con datos de secuencias genómicas, farmacología de drogas y documentos textuales. Los modelos LDA que emplean RPC produjeron matrices útiles, demostrando su aplicabilidad a diversos tipos de datos.
- En su estudio, (Sbalchiero & Eder, 2020) analizan la relación entre la longitud del texto y el número ideal de temas mediante una relación de ley de potencia y un modelo matemático. Los resultados muestran que el número óptimo de temas, determinado por el método de verosimilitud logarítmica, se mantiene constante inicialmente, pero disminuye para segmentos de texto de 20,000 a 50,000 palabras, especialmente cuando superan las 10,000 palabras. Esto sugiere una correlación entre el tamaño del texto y el número óptimo de temas, proporcionando una base para futuras investigaciones en textos de gran longitud.
- En su estudio, (Sbalchiero & Eder, 2020) analizan la relación entre la longitud del texto y el número ideal de temas, utilizando una ley de potencia y un modelo matemático. Los resultados muestran que el número óptimo de temas, determinado mediante verosimilitud logarítmica, se mantiene constante en textos más cortos, pero tiende a disminuir cuando la longitud del segmento aumenta, especialmente en el rango de 20,000 a 50,000 palabras. Esta observación sugiere una correlación entre el tamaño del texto y la complejidad temática, lo que proporciona una base para futuras investigaciones en el análisis de textos extensos.
- En su estudio (Dang & Nguyen, 2018), presentan ComModeler, una técnica para modelar temas mediante la identificación de comunidades en redes dinámicas. El proceso comienza extrayendo y organizando términos clave en una red. Luego, analiza la frecuencia de términos, los cambios abruptos y la centralidad de los vértices para descubrir comunidades. Esta técnica utiliza las propiedades de la red para trazar temas históricos, identificar marcas de tiempo y conectar términos co-ubicados, ofreciendo un enfoque más interactivo y detallado para la modelación de temas.

4. Metodología

Este capítulo describe el enfoque metodológico empleado para realizar un análisis de las descripciones asociadas a las solicitudes de información pública gubernamental a través de la Plataforma Nacional de Transparencia (PNT). Para una mejor comprensión del proceso metodológico utilizado, la Figura 4.1ilustra las etapas clave y la secuencia seguida en este estudio.

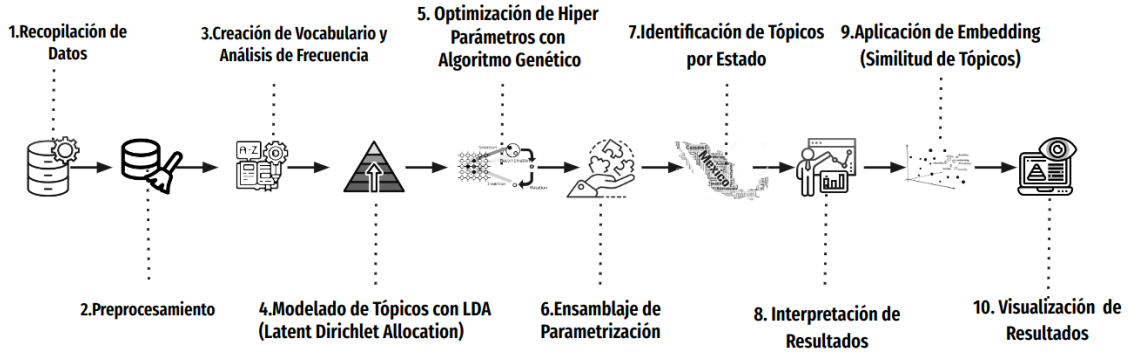

Figura 4.1. Proceso metodológico.

4.1 Recopilación de información:

La información analizada fue obtenida de la sección de datos abiertos de la PNT, enfocada en las entidades federales sujetas a obligaciones de transparencia. Por consiguiente, se excluyeron las solicitudes dirigidas a entidades no federales, como gobiernos estatales y locales, órganos legislativos y judiciales, partidos políticos, sindicatos y otras organizaciones similares. Los datos están disponibles en formatos como SQL, XML, CSV y JSON.

Para el análisis se utilizaron datos del periodo 2003–2015, extraídos de Infomex, una versión precedente de la PNT. A partir de 2015, la información se obtuvo directamente de dicha plataforma (véase la Figura 4.2). A lo largo del periodo examinado, se analizaron aproximadamente 2,518,875 solicitudes de información pública dirigidas a diversas dependencias federales.

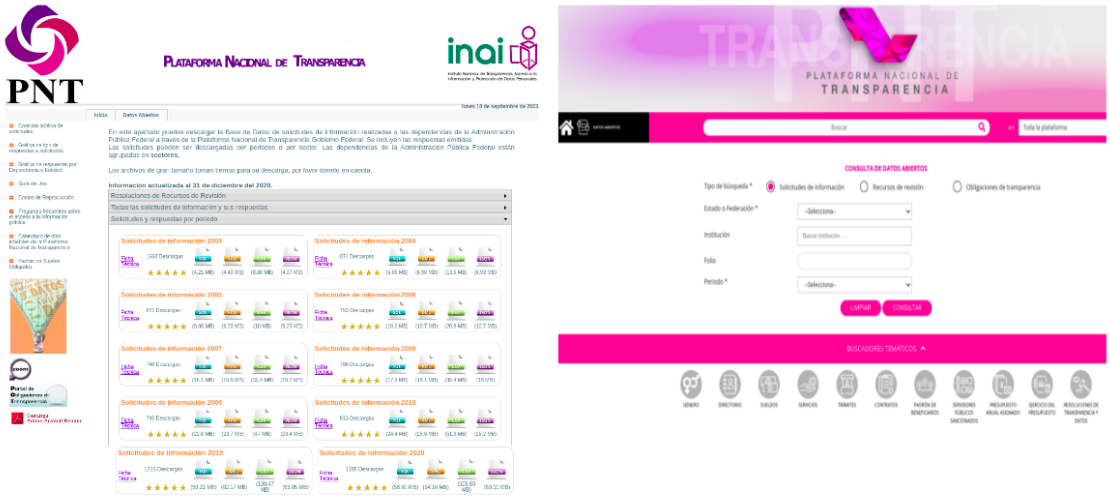

Figura 4.2. Plataforma Infomex y Plataforma Nacional de Transparencia.

Para los propósitos de esta investigación, se descargaron los datos en formato JSON desde la PNT, y posteriormente se clasificaron según los estados de la República Mexicana para facilitar un análisis detallado por entidad federativa y año. Es importante destacar que, a partir del año 2016, algunos registros omitieron la especificación del estado como medida de protección de la privacidad personal. En estos casos, las solicitudes fueron conservadas en el análisis general, pero no se incluyeron en la desagregación por entidad federativa. La base de datos contiene detalles completos sobre las consultas de ciudadanos enviadas a diversas entidades gubernamentales, así como sus respectivas respuestas. En total, se incluyen alrededor de 20 atributos, relativos a cada solicitud.

La Tabla 4.1 muestra los atributos de una solicitud de información relacionada con el COVID-19, con base en datos de la Plataforma Nacional de Transparencia (Instituto Nacional de Transparencia, Acceso a la Información y Protección de Datos Personales (INAI), 2023). Dicha solicitud busca datos sobre la disponibilidad de pruebas gratuitas, los tratamientos existentes y sus costos, así como los procedimientos para acceder a atención médica en Mérida, Yucatán, México. Este ejemplo subraya la complejidad de la información no estructurada, resaltando la importancia de entender las necesidades y consultas específicas de los ciudadanos.

Tabla 4.1. Atributos de la base de datos de solicitudes del INAI.

Atributo INAI	**Descripción**	**Ejemplo**
Folio	Identificador único de solicitud de 13 dígitos.	310572322000377
Fecha de solicitud	Fecha y hora en que se realizó la solicitud.	05/07/2022, 14:23
Dependencia	Nombre del departamento gubernamental al que se envía la solicitud.	Servicios de salud de Yucatán
Estado	Estado actual de la solicitud, p.ej., en proceso.	Terminada
Medio de entrada	Cómo se presentó la solicitud: electrónica o manual.	Electrónica
Tipo de solicitud	Naturaleza de la solicitud: información pública, datos personales, corrección de datos.	Información pública
Descripción	Explicación del solicitante sobre la información requerida.	Detalles de pruebas y tratamientos de COVID-19 en Mérida, Yucatán.
Otros datos	Detalles adicionales para facilitar la localización de la información.	N/A
Archivo adjunto de solicitud	URL de archivos suplementarios proporcionados por el solicitante.	N/A
Medio de entrega	Método preferido para recibir la información solicitada.	Electrónico a través del sistema PNT
Fecha límite de respuesta	Fecha límite para que el departamento responda.	20/07/2022
Respuesta	Tipo de respuesta proporcionada por el departamento gubernamental.	Entrega de información vía PNT
Archivo de respuesta	Archivo digital proporcionado por el departamento como parte de la respuesta.	N/A
Fecha de respuesta	Fecha en que el departamento respondió.	14/07/2022
País	País de ubicación del solicitante.	México
Estado	Estado de ubicación del solicitante dentro del país.	Yucatán
Municipio	Municipio de ubicación del solicitante.	Mérida
Código Postal	Código postal de ubicación del solicitante.	97098
Sector	Sector gubernamental del departamento al que se dirige.	Descentralizados

Fuente: *Elaboración propia con datos del INAI (2023).*

La Figura 4.3, muestra un ejemplo de la descripción de una solicitud de información relacionada con el COVID-19, en la que un ciudadano requiere información sobre aspectos críticos como la disponibilidad de pruebas gratuitas, la existencia de tratamientos, sus costos asociados, y los procedimientos específicos para recibir atención médica en Mérida, Yucatán. Esta petición refleja la necesidad de información vital en medio de una crisis sanitaria, y la urgencia y ansiedad de los ciudadanos por acceder a servicios de salud eficientes que puedan gestionar adecuadamente la pandemia.

Las preguntas reflejan la preocupación por entender la interacción con el sistema de salud, resaltando la demanda de transparencia y agilidad en las respuestas por parte de las autoridades gubernamentales. Esta solicitud evidencia la importancia de una comunicación clara y accesible durante emergencias de salud pública. Las respuestas a estas preguntas son importantes tanto para el solicitante como para la comunidad en general, ya que proporcionan información valiosa que puede ser útil para otras personas en situaciones similares.

ESTATU	↓ MEDI	TIPOSC	DESCRIPCIONSOLICITUD	OTROS	ARCHIVOA
Termin	Electró	Inform	Con relación al COVID 19, solicito la siguiente información: ¿Dónde se pueden aplicar las pruebas gratuitas? En caso de dar positivo, ¿Existe algún tratamiento? ¿Tiene algún costo dicho tratamiento? ¿Cuál es el procedimiento para tener acceso al calendario y los lugares donde se aplican las pruebas gratuitas de COVID19 en la ciudad de Mérida Yucatán ¿Cuál es el precio para acceder a un tratamiento en caso de dar positivo por COVID19 ? ¿Cuáles son los requisitos para acceder a un tratamiento en caso de dar positivo por COVID19 en los hospitales públicos?	None	https://ser
Termin	Electró	Inform		None	https://ser
Termin	Electró	Inform		None	https://ser
Termin	Electró	Inform		None	https://ser
Termin	Electró	Inform		None	https://ser
Termin	Electró	Informa	En relación con la capacidad de la Institución para recibir nuevos alumnos en todas sus e	None	https://ser

Figura 4.3. Ejemplo de descripción de solicitud.

4.2 Creación de archivos JSON

Después de la obtención de datos, se procede a la creación de archivos en formato JSON, estructurados según la entidad federativa y el año correspondiente. Este proceso se realiza utilizando como referencia la columna `estado` de nuestra base de datos para la clasificación de la información.

El siguiente ejemplo muestra la estructura del archivo para el estado de Aguascalientes en el año 2003, denominado `Aguascalientes_2003.json`:


```
{
"SOLICITUDES": {
"SOLICITUD": [
{
"Folio": "0002100000103",
"FechaSolicitud": "2003/06/12 08:27:44",
"Dependencia": "SECRETARÍA DE TURISMO",
"Estatus": "Terminada",
"MedioEntrada": "Electrónica",
"TipoSolicitud": "Información Pública",
"DescripcionSolicitud": "Solicito información sobre trabajadores...",
"OtrosDatos": "",
"ArchivoAdjuntoSolicitud": "",
"MedioEntrega": "Entrega por Internet en el INFOMEX",
"FechaLimite": "2003/07/10",
"Respuesta": "Entrega de información en medio electrónico",
"TextoRespuesta": "ANEXO AL PRESENTE ENCONTRARA EL ARCHIVO EN PDF...",
"ArchivoRespuesta": "https://www.infomex.org.mx...",
"FechaRespuesta": "2003/06/25",
"Pais": "México",
"Estado": "Aguascalientes",
"Municipio": "AGUASCALIENTES",
"CodigoPostal": "20190",
"Sector": "Turismo"
 }
 ]
 }
}
```

4.3 Pre-procesamiento de las solicitudes de información del INAI

El proceso de preparación de las solicitudes de información es importante en el manejo de datos, ya que implica una minuciosa limpieza, transformación y organización de la información. Esta etapa es fundamental para optimizar los datos para análisis y modelado futuros.

4.3.1 Limpieza de datos

En primer lugar, se realiza una limpieza de datos, la cual incluye la eliminación de caracteres especiales como símbolos [, #, %], números y signos de puntuación [*punto, coma, punto y coma*], con la finalidad de facilitar el procesamiento subsecuente.

Por ejemplo, considerando el texto de solicitud relacionado con el COVID-19, es necesario eliminar números y caracteres especiales para centrarse en el contenido textual que requiere análisis. En este caso específico, la frase *Con relación al COVID 19, solicito la siguiente información:* se limpiaría para eliminar "19", dejando *Con relación al COVID, solicito la siguiente información*. Esta simplificación ayuda a estandarizar las menciones de enfermedades o términos clave, lo que es vital para agrupar y comparar solicitudes similares dentro de grandes conjuntos de datos.

Además, al depurar el texto de signos de puntuación y separadores, como en las preguntas *¿Dónde se pueden aplicar las pruebas gratuitas?* o *¿Cuál es el procedimiento para tener acceso al calendario y los lugares donde se aplican las pruebas gratuitas de COVID-19 en la ciudad de Mérida, Yucatán?*, se mejora la detección de frases clave y se facilita la segmentación del texto en componentes relevantes para análisis más detallados.

4.3.2 Tokenización

La tokenización es un paso esencial en el análisis de textos, especialmente pertinente en el campo del procesamiento de lenguaje natural (PLN). Este proceso consiste en descomponer el texto en unidades básicas llamadas tokens, que son esencialmente segmentos o palabras individuales. Para ilustrar el proceso de tokenización, consideremos un ejemplo práctico con el texto de la descripción de una solicitud:

Con relación al COVID 19, solicito la siguiente información: - ¿Dónde se pueden aplicar las pruebas gratuitas? - En caso de dar positivo, ¿Existe algún tratamiento? ¿Tiene algún costo dicho tratamiento? - ¿Cuál es el procedimiento para tener acceso al calendario y los lugares donde se aplican las pruebas gratuitas de COVID19 en la ciudad de Mérida Yucatán? - ¿Cuál es el precio para acceder a un tratamiento en caso de dar positivo por COVID19? - ¿Cuáles son los requisitos para acceder a un tratamiento en caso de dar positivo por COVID19 en los hospitales públicos?

A partir de este texto, el proceso de tokenización se realizaría descomponiendo el texto en los siguientes tokens: “con”, “relación”, “al”, “COVID”, “solicito”, “la”, “siguiente”, “información”, “¿Dónde”, “se”, “pueden”, “aplicar”, “las”, “pruebas”, “gratuitas”, “¿En”, “caso”, “de”, “dar”, “positivo”, “¿Existe”, “algún”, “tratamiento”, “¿Tiene”, “algún”, “costo”, “dicho”, “tratamiento”, “¿Cuál”, “es”, “el”, “procedimiento”, “para”, “tener”, “acceso”, “al”, “calendario”, “y”, “los”, “lugares”, “donde”, “se”, “aplican”, “las”, “pruebas”, “gratuitas”, “de”, “COVID”, “en”, “la”, “ciudad”, “de”, “Mérida”, “Yucatán”, “¿Cuál”, “es”, “el”, “precio”, “para”, “acceder”, “a”, “un”, “tratamiento”, “en”, “caso”, “de”, “dar”, “positivo”, “por”, “COVID”, “¿Cuáles”, “son”, “los”, “requisitos”, “para”, “acceder”, “a”, “un”, “tratamiento”, “en”, “caso”, “de”, “dar”, “positivo”, “por”, “COVID”, “en”, “los”, “hospitales”, “públicos”.

4.3.3 Eliminación de palabras vacías

La eliminación de palabras vacías es esencial para concentrar el análisis en los términos más informativos del texto. Utilizamos un filtro personalizado que excluye palabras comunes, preposiciones, términos de saludo, términos legales y específicos de solicitudes de información. La Tabla 4.2 muestra ejemplos de estas categorías de palabras vacías:

Tabla 4.2. Ejemplos de palabras vacías excluidas en el pre-procesamiento.

Categoría	**Ejemplos**
Términos comunes	“de”, “y”, “en”, “que”, “la”, “el”, “los”, “las”
Preposiciones	“sobre”, “ante”, “bajo”, “con”, “contra”, “desde”, “durante”
Términos de saludo	“hola”, “buenos días”, “atentamente”, “cordiales saludos”, “estimado/a”
Términos legales	“de acuerdo con”, “según lo establecido por”, “con base en”, “en virtud de”, “conforme a”
Términos de solicitud	“solicitud”, “requerimiento”, “petición”, “solicito”, “requiero”, “solicitado”

4.4 Creación de vocabulario

Durante esta etapa del proceso, se desarrolló un vocabulario a partir de las palabras y n-gramas encontrados en las solicitudes de información pública. Inicialmente, se generó un diccionario de frecuencias para cada token , y se extendió este vocabulario básico incluyendo n-gramas de tamaño tres, que concatenan tres términos consecutivos para capturar combinaciones de palabras que aparecen juntas frecuentemente.

Se analizó la frecuencia de aparición de cada término y n-grama en el conjunto de datos y se estudió la distribución de estas frecuencias utilizando la ley de Zipf. Esta ley permite comprender cómo se distribuyen las frecuencias de las palabras en un texto, llevando a eliminar aquellas de muy baja frecuencia y a categorizar los términos menos usuales como palabra desconocida. Este enfoque reduce eficazmente el tamaño del vocabulario, optimizando la eficiencia del modelo sin comprometer su precisión o alcance.

Este proceso involucra varios pasos: cargar los datos desde un archivo CSV, calcular las frecuencias de las palabras y n-gramas, aplicar una transformación logarítmica a estas frecuencias para estabilizar la dispersión de datos y resaltar características significativas, y luego organizarlas de mayor a menor en un ranking. Este ranking, junto con el logaritmo de las frecuencias (log_f), ayuda a ajustar la distribución para un análisis más accesible. La Tabla 4.3 refleja cómo los términos más comunes, como *los*, *para*, *información*, *número*, dominan en frecuencia, mientras que términos más específicos como *solicita* o *tipo* aparecen con menor frecuencia, lo que destaca la diversidad y especificidad del vocabulario utilizado.

Tabla 4.3. Ejemplo de vocabulario optimizado.

Frecuencia	**Palabra**	**Log_f**	**Rango**
676	información	6.516	1
663	número	6.497	2
407	informe	6.009	3
343	medio	5.838	4
341	copia	5.832	5
277	nombre	5.624	6
260	así	5.561	7
229	saber	5.434	8
202	instituto	5.308	9
196	solicita	5.278	10
190	tipo	5.247	11

A continuación, elaboramos un gráfico (véase Figura 4.4) que muestra la relación entre el rango de las palabras y su frecuencia logarítmica. Este gráfico es esencial para visualizar la distribución de las frecuencias y determinar dónde se encuentra el knee point, indicando el tamaño óptimo del vocabulario. Utilizamos KneeLocator, una herramienta que implementa un algoritmo específico para identificar este punto crítico, que se reconoce por un cambio en la pendiente de la curva, señalando una disminución significativa en la tasa de frecuencia de las palabras.

Identificar el punto de inflexión nos permite establecer el tamaño óptimo del vocabulario, representando el conjunto de palabras que más influyen en la variabilidad de los datos antes de que la frecuencia de aparición disminuya de manera abrupta. Este criterio ayuda a distinguir entre términos comunes y raros, asegurando un análisis de datos efectivo que mantiene tanto la integridad como la relevancia del estudio de las solicitudes de información pública.

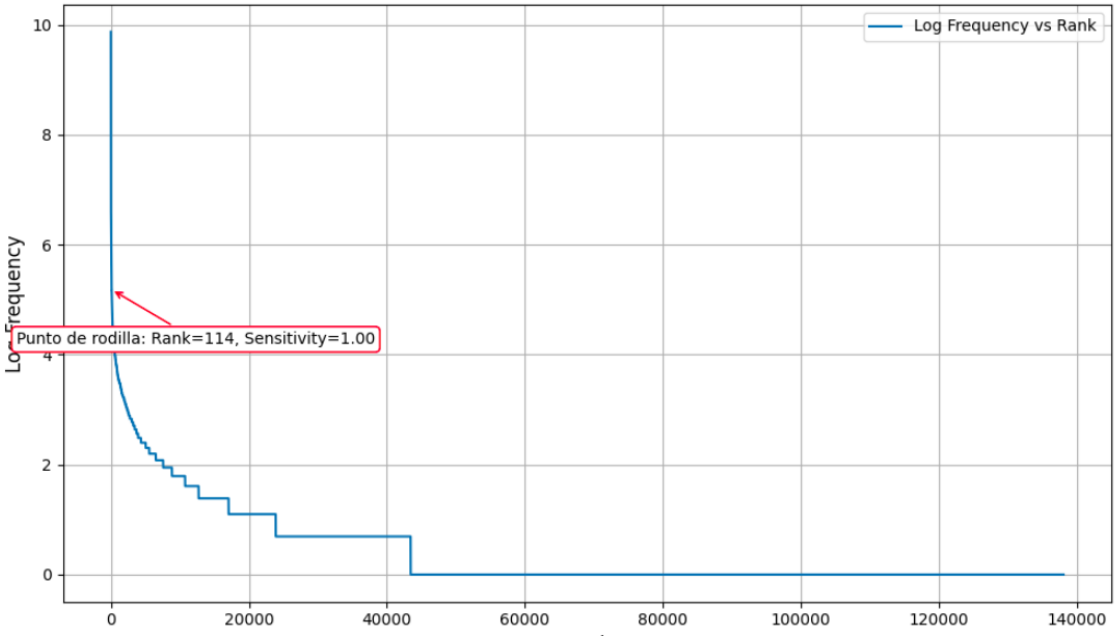

Figura 4.4. Optimización de vocabulario en solicitudes de Jalisco (2020).

Posteriormente, para seleccionar solo aquellas palabras cuyo rango es igual o inferior al punto de rodilla, se exporta un archivo de texto con el subconjunto filtrado. Además, se genera una versión del vocabulario reducido en formato texto, donde cada palabra se delimita con comillas y se separa por comas, lo cual permite su uso directo en la siguiente etapa del procesamiento.

4.5 Modelado de tópicos con LDA

En esta fase se implementa el modelo de LDA, con el objetivo de generar una serie de archivos claves para el avance del proyecto. Estos archivos incluyen Corpus.mm, que representa una colección estructurada de los datos textuales; Diccionario.dict, un archivo que mapea cada palabra única a un identificador numérico; Palabras Descartadas.txt, que contiene un listado de términos filtrados durante la fase de limpieza de datos; títulos.txt, donde se almacenan los títulos o encabezados asociados a cada solicitud de información; y vocabulario.txt, que es un compendio de todas las palabras relevantes identificadas en el corpus.

La fórmula (4.1) es utilizada en el contexto de modelos de temas, como el LDA y otros modelos de mezcla de temas. Estos modelos son técnicas de modelado estadístico que se utilizan para describir cómo se generan los documentos en una colección en términos de una mezcla de varios temas. Donde, un tema es una distribución sobre un vocabulario fijo, y cada documento se considera una mezcla de varios temas. La fórmula tiene el propósito de calcular la probabilidad de observar una palabra específica.

$$P(w_i) = \sum_{j=1}^{T} P(w_i \mid z_i = j) \cdot P(z_i = j) \tag{4.1}$$

- $P(w_i)$: Probabilidad marginal de la palabra w_i.
- $\sum$: Sumatoria sobre todas las categorías j de 1 a T.
- $P(w_i \mid z_i = j)$: Probabilidad de w_i dado el tema j.
- $P(z_i = j)$: Probabilidad a priori del tema j.

A continuación, se describe brevemente el funcionamiento del modelo LDA. Este modelo analiza las palabras contenidas en un conjunto de documentos en este caso, solicitudes realizadas a la (PNT) y las agrupa en tópicos latentes con base en su distribución estadística. Cada tópico se representa como una distribución de palabras, y cada documento como una combinación ponderada de estos tópicos.

Como ejemplo, se presenta una solicitud dirigida a la PNT, en la cual se destacan en **negritas** las palabras clave que el modelo LDA podría asociar con distintos tópicos.

Ejemplo de solicitud:

*Solicito información sobre cómo sus programas de **Nutrición**, **Ejercicio** y **Bienestar** abordan la **Prevención** de enfermedades y promueven la **Salud Mental**, considerando el **impacto ambiental**. Me interesa conocer las estrategias para gestionar los **Gastos en Salud**, reducir la **Contaminación** y minimizar la **Huella Ecológica** mediante prácticas de **Reciclaje** de residuos de salud. Además, agradecería detalles sobre las prácticas de **Finanzas** y **Planificación Financiera**, cuentas en inversiones relacionadas con **Energías Renovables** y proyectos de **Conservación ambiental**.*

La Tabla 4.4 muestra un ejemplo de categorías de tópicos derivados de las palabras clave contenidas en la solicitud.

Tabla 4.4. Ejemplo de categorías de tópicos derivadas de una solicitud.

Salud	**Medio Ambiente**	**Presupuesto**
Nutrición	Reciclaje	Finanzas
Ejercicio	Contaminación	Gasto
Bienestar	Residuos de salud	Gestión
Prevención	Energías	Inversiones
Enfermedad	Conservación ambiental	Planificación
Salud mental	Renovables	Financiera
Gasto	Huella ecológica	Cuentas
		Contabilidad
		Costos

A partir del análisis realizado con LDA, es posible identificar agrupaciones temáticas como las siguientes:

- En el tema *Medio Ambiente*, palabras como *reciclaje*, *contaminación* y *renovables* son comúnmente asociadas por LDA con este tópico, debido a su frecuencia y relevancia en solicitudes relacionadas con la gestión ambiental.
- Para el tema *Salud*, términos como *nutrición*, *bienestar*, y *prevención* son clave para identificar solicitudes que traten sobre programas de salud pública.
- En el área de *Presupuesto*, palabras como *finanzas*, *gasto* y *inversiones* ayudan a LDA a clasificar solicitudes enfocadas en la administración financiera y presupuestaria.

Este comportamiento se basa en los principios probabilísticos del modelo:

- $P(w_i|z_i = j)$: Probabilidad de que una palabra w_i esté significativamente asociada con un tópico $z_i = j$. Por ejemplo, la probabilidad de que *reciclaje* esté asociada con el tópico de *Medio Ambiente* es alta.
- $P(z_i = j)$: Indica la prevalencia del tópico z_i en un documento. Si una solicitud contiene numerosas menciones de términos relacionados con *energías renovables*, la probabilidad de que ese tópico domine en el documento es considerable.

4.6 Evaluación de tópicos

En el modelado de tópicos, la evaluación de la coherencia comienza con la identificación de palabras clave, seleccionadas basándose en su frecuencia y pertinencia temática. La medida C_v, también conocida como “Coherence Score” (Puntuación de Coherencia), es una métrica utilizada en el modelado de tópicos para evaluar la coherencia de los tópicos identificados dentro de un conjunto de datos o corpus.La medida de coherencia con mejor rendimiento, C_V, combina la similitud de coseno indirecta con la medida de Normalized Pointwise Mutual Information (NPMI) y una ventana deslizante booleana (Röder et al., 2015).

La fórmula Información NPMI 4.2 es una medida estadística de asociación de palabras utilizada en el procesamiento de lenguaje natural y en el modelado de tópicos

$$\text{NPMI}(w', w^*) = \frac{\log\left(\frac{P(w',w^*)+\epsilon}{P(w')P(w^*)+\epsilon}\right)}{-\log(P(w', w^*) + \epsilon)} \tag{4.2}$$

Donde:

- $P(w', w^*)$: La probabilidad de que las palabras w' y w^* co-ocurran.
- $P(w')$: La probabilidad de que la palabra w' ocurra en el corpus.
- $P(w^*)$: La probabilidad de que la palabra w^* ocurra en el corpus.
- ϵ: Una constante pequeña para evitar la división por cero y para manejar casos donde las palabras no co-ocurren.

Los valores de NPMI varían entre -1 y 1, donde 1 indica que las palabras siempre co-ocurren, 0 significa que las palabras co-ocurren con la frecuencia que se esperaría por azar, y -1 significa que las palabras nunca co-ocurren.

Para calcularlo, se emplea la medida NPMI, que cuantifica qué tan semánticamente relacionadas están las parejas de palabras dentro de un tópico. El puntaje C_V agrega los valores NPMI de múltiples pares de palabras, proporcionando una estimación de la coherencia general del tópico. Un valor más alto de C_V indica una mayor coherencia, es decir, que las palabras dentro del tópico están más estrechamente relacionadas entre sí.

4.7 Ejemplo de cálculo de coherencia c_v

A continuación, se presenta un ejemplo basado en (Rijcken, 2023). Consideremos un corpus compuesto por cuatro documentos, denotado por un total de $D = 4$ elementos:

Consideremos un corpus de cuatro documentos:

$$C = \{\delta_1, \delta_2, \delta_3, \delta_4\}$$

donde:

$$\delta_1 = \{\text{presupuesto}, \text{salud}, \text{municipal}\}, \qquad \delta_2 = \{\text{presupuesto}, \text{educación}, \text{estatal}\}$$
$$\delta_3 = \{\text{contratos}, \text{licitación}, \text{transparente}\}, \qquad \delta_4 = \{\text{gasto}, \text{publicidad}, \text{gubernamental}\}$$

Las categorías más probables son:

$$W_1 = \{\text{salud}, \text{municipal}\}, \quad W_2 = \{\text{educación}, \text{estatal}\}, \quad W_3 = \{\text{licitación}, \text{transparente}\}$$

Con un tamaño de ventana de 3, las probabilidades relevantes son:

$$P(\text{licitación}) = \tfrac{1}{5}, \quad P(\text{presupuesto}) = \tfrac{2}{5}, \quad P(\text{presupuesto}, \text{educación}) = \tfrac{1}{5}$$

Puntajes de NPMI:

$$\text{NPMI}(\text{licitación}, \text{transparente}) = 1{,}113, \quad \text{NPMI}(\text{presupuesto}, \text{educación}) = 0{,}571$$

Finalmente, el puntaje c_v es: $c_v = 0{,}838$

4.8 Desarrollo e implementación de algoritmo genético

En esta fase se desarrolló una versión específica del algoritmo genético para la optimización eficiente de los parámetros α y β y K, utilizando como función fitness la coherencia. La Figura 4.5 proporciona un diagrama de flujo que describe esta optimización. El algoritmo genético emplea una configuración particular que le permite explorar el espacio de busqueda: un tamaño de población de 20 para diversidad, 50 generaciones para asegurar una evolución adecuada, selección de 4 padres para una mezcla genética equilibrada, retención de 2 soluciones élite para preservar genes superiores y una tasa de mutación del 0.05 para fomentar nuevos rasgos. Además, explora un espacio genético que incluye los parámetros α y β dentro de un rango de 0.01 a 1.0, y K entre 2 y 50, asegurando que el algoritmo ajuste los parámetros del modelo e identifique los temas más coherentes para el análisis. La efectividad del algoritmo está respaldada por pruebas estadísticas, cuyos detalles se proporcionan en el trabajo referenciado (Kuri-Morales et al., 2013).

4.8.1 Inicializando la población.

La función initialize_population se encarga de generar una población inicial, cuyos detalles se especifican en la Tabla 4.5. Esta función acepta dos parámetros: pop_size, que determina el número total de individuos a generar, y parameter_ranges, que define los límites dentro de los cuales los valores de los parámetros deben ser seleccionados.

La población se inicializa como una lista vacía, y mediante un bucle que se ejecuta pop_size veces, se generan individuos de manera iterativa. Cada individuo se compone de valores asignados aleatoriamente dentro de los límites especificados en parameter_ranges, empleando para ello la función random.uniform. Este procedimiento garantiza que cada valor de parámetro se mantenga dentro de los rangos permitidos. Específicamente, el parámetro y K requiere un tratamiento especial, ya que debe ser un entero; por tanto, el último valor en la lista de cada individuo se convierte a este tipo de dato.

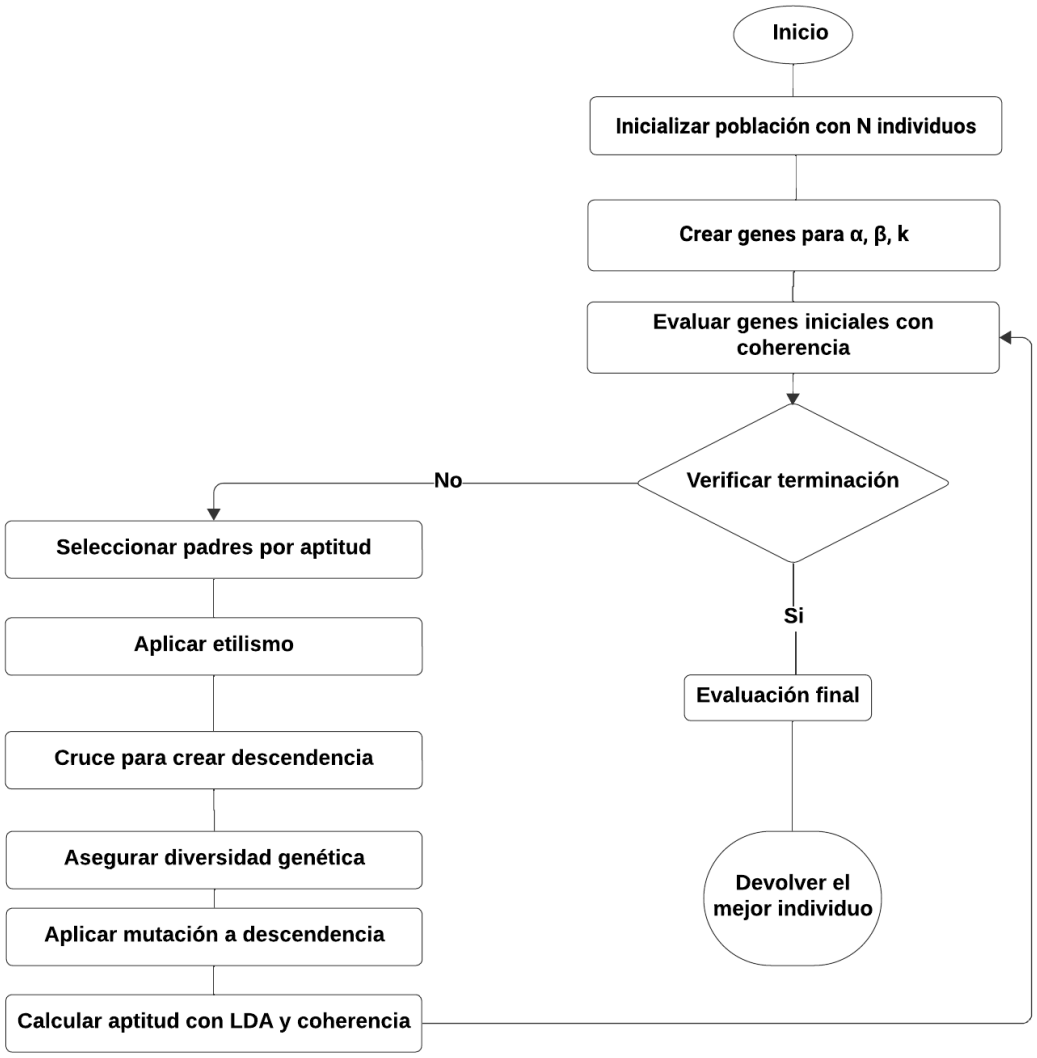

Figura 4.5. Diagrama de flujo del algoritmo genético.

Tabla 4.5. Identificadores para la función de inicialización de la población.

Identificador	Definición
`pop_size`	Tamaño total de la población a inicializar.
`parameter_ranges`	Diccionario que mapea nombres de parámetros a tuplas, cada una especificando los límites mínimos y máximos admisibles para ese parámetro.
`population`	Lista que almacena los individuos de la población, donde cada individuo se representa mediante una lista de valores de parámetros.
`individual`	Lista de valores que representa a un solo individuo, cada valor es generado dentro de los rangos establecidos; el último valor se transforma en entero para ajustarse a la especificación del parámetro K.

4.8.2 Función de ajuste de coherencia

La función evaluate_coherence se implementa para medir la coherencia de los temas generados por un modelo de LDA, un indicador crítico en la evaluación de la utilidad de modelos de temas según (Stevens et al., 2012) y (Omar et al., 2015). Esta función verifica que los temas identificados sean interpretables y cohesivos, aspectos fundamentales para confirmar su pertinencia en contextos aplicados.

La función opera en el entorno del módulo LdaModel de Gensim, donde el modelo LDA se construye y optimiza a través de iteraciones sucesivas sobre el corpus de datos. El número de iteraciones es un determinante para la fidelidad y precisión del modelo resultante. Para la evaluación de la coherencia, se emplea la clase CoherenceModel de Gensim, configurando un objeto coherence_model_lda que facilita la evaluación de la coherencia temática.

Adicionalmente, se integra una función de ajuste que cuantifica y retorna los valores de coherencia para cada modelo en una población de modelos evaluados. Esta función inicia generando una lista para almacenar las puntuaciones de coherencia de cada modelo, resaltando así la capacidad de cada modelo para producir temas coherentes y significativos. La Tabla 4.6 proporciona las definiciones de los identificadores utilizados en la evaluación de la coherencia:

Tabla 4.6. Identificadores y sus definiciones en la función de evaluación de coherencia.

Identificador	Definición
α	Hiperparámetro que regula la distribución de temas en los documentos. Un valor bajo promueve una mayor exclusividad de temas en cada documento.
β	Hiperparámetro que gestiona la distribución de palabras dentro de los temas. Un valor bajo favorece una mayor concentración de ciertas palabras clave en cada tema.
k	Cantidad de temas que el modelo debe extraer del corpus.
`corpus`	Colección de documentos transformada en una estructura de datos que lista frecuencias de palabras por documento.
`id2word`	Diccionario que vincula identificadores numéricos con sus palabras correspondientes.
`texts`	Conjunto de textos originales que han sido preprocesados para análisis.
`lda_model`	Modelo LDA configurado y optimizado para analizar el corpus proporcionado.
`coherence_model_lda`	Modelo dedicado a evaluar la coherencia de los temas inferidos por el modelo LDA, asegurando su correspondencia con el contenido de los textos.

4.8.3 Selección de padres

La selección de padres constituye un mecanismo fundamental en los algoritmos genéticos, al priorizar a los individuos más aptos para la reproducción. Este proceso está diseñado para favorecer la transmisión de características ventajosas a las generaciones futuras, bajo la premisa de que los individuos con mayor aptitud tienen una mayor probabilidad de heredar sus atributos deseables a sus descendientes.

Para la implementación de este componente, se utiliza la biblioteca NumPy, que proporciona una gestión de arrays numéricos. Mediante el uso de la función np.argsort() aplicada al array fitness, se clasifican los individuos en función de su nivel de aptitud. A continuación, los num_-parents mejores clasificados se seleccionan para participar en el proceso reproductivo. La Tabla 4.7 detalla los identificadores empleados en el proceso de selección de padres.

Tabla 4.7. Identificadores para la función de selección de padres.

Identificador	Definición
`population`	Conjunto de todos los individuos en la generación actual, de los cuales se seleccionará un subconjunto como padres.
`fitness`	Array que contiene los valores de aptitud para cada individuo en la población, utilizado para determinar los más idóneos para la reproducción.
`num_parents`	Número de individuos a seleccionar como padres para la siguiente generación, basado en su superior aptitud.
`parents`	Subconjunto de individuos seleccionados para la reproducción, identificados por poseer los mayores valores de aptitud dentro de la población.

4.8.4 Cruce genético

El cruce genético es un componente esencial en los algoritmos genéticos, que facilita la combinación de características genéticas de dos progenitores para producir descendencia. Este proceso es fundamental para la exploración y explotación del espacio de soluciones, permitiendo mejoras en la aptitud de las futuras generaciones. Se implementa una técnica de cruce en un punto específico del vector de parámetros de la descendencia, como se detalla en la Tabla 4.8, para distribuir equitativamente las contribuciones genéticas de ambos padres.

La operación de cruce comienza con la iteración de un bucle que corresponde al número deseado de descendientes a generar. En cada iteración, se selecciona un par de padres basado en índices cíclicos para asegurar una distribución equitativa y diversa de material genético. La descendencia resultante se produce mediante la concatenación de segmentos genéticos de los progenitores, seleccionados antes y después de un punto de cruce calculado. Este punto de cruce se establece típicamente en la mitad de la longitud del vector de parámetros, permitiendo así una fusión eficaz de los genes y fomentando la herencia de rasgos beneficiosos.

4.8.5 Mutación

La mutación es un mecanismo esencial en los algoritmos genéticos, diseñado para aumentar la diversidad genética dentro de una población y evitar la convergencia prematura hacia óptimos locales subóptimos. Este proceso introduce variaciones genéticas en los individuos de manera controlada, permitiendo la exploración de nuevos espacios de soluciones que podrían pasar desapercibidos de otro modo.

Tabla 4.8. Identificadores para la función de cruce genético.

Identificador	Definición
`parents`	Array que almacena la información genética de los progenitores seleccionados para el cruce, de cuyos genes se derivará la descendencia.
`offspring_size`	Tupla que determina el número de descendientes a generar (`offspring_size[0]`) y la dimensionalidad de cada descendiente (`offspring_size[1]`), definiendo así el tamaño y la estructura del array de descendencia resultante.
`offspring`	Array de descendencia generado por el proceso de cruce, donde cada individuo hereda características genéticas de dos progenitores distintos.
`crossover_point`	Punto calculado que divide la información genética de los progenitores, ubicado típicamente a la mitad de la segunda dimensión del vector de parámetros de cada descendiente.
`parent1_idx`, `parent2_idx`	Índices utilizados para seleccionar los pares de progenitores para la combinación genética, facilitando una diversidad genética adecuada en la descendencia.

El proceso de mutación se rige por una tasa de mutación específica, que establece la probabilidad con la que cada individuo en la población puede experimentar cambios en sus parámetros genéticos. La operación de mutación se inicia evaluando cada descendiente contra esta tasa: se genera un número aleatorio entre 0 y 1 y, si este número es inferior a la tasa de mutación, se procede a la mutación del individuo.

En la práctica, se selecciona al azar un parámetro del conjunto de parámetros susceptibles de mutación para ser alterado, tal como se documenta en la Tabla 4.9. Para parámetros con requisitos específicos, como el K, que debe ser un entero, se asigna un nuevo valor dentro de un rango predefinido de forma aleatoria. Esto garantiza la introducción de variabilidad genética adecuada y el cumplimiento de las restricciones de valores de parámetros válidos, potenciando así la capacidad del algoritmo genético para explorar y descubrir soluciones.

Tabla 4.9. Identificadores para la función de mutación en un algoritmo genético.

Identificador	Definición
`offspring`	Conjunto de individuos de la generación actual sujetos a mutación.
`parameter_ranges`	Diccionario que delimita los rangos permitidos para cada parámetro susceptible a mutación, asegurando que las mutaciones se mantengan dentro de límites válidos.
`mutation_rate`	Probabilidad de que cualquier individuo de la población experimente una mutación, dictando la frecuencia de este evento dentro de la población.
`param_to_mutate`	Parámetro específico seleccionado al azar para su mutación, determinado a partir de las claves del diccionario `parameter_ranges`.
`index`	Posición del parámetro a mutar dentro de la secuencia de parámetros del individuo, necesario para aplicar la mutación a parámetros distintos de K.

4.8.6 Diversidad genética

La diversidad genética permite prevenir la convergencia prematura hacia óptimos locales y mejora la habilidad del algoritmo para explorar el espacio de soluciones de manera efectiva. Este principio evalúa la contribución de cada individuo a la varianza genética de la población, comparando su aptitud con un umbral predeterminado. Un valor de aptitud que se sitúa por debajo de este umbral sugiere una insuficiente contribución a la diversidad de la población, lo que puede ser indicativo de una estagnación evolutiva.

En respuesta a tales circunstancias, se implementa un proceso de regeneración para los individuos afectados. Este procedimiento involucra la reasignación de parámetros genéticos, generando nuevos valores aleatorios dentro de los rangos permitidos especificados, asegurando la conformidad con los límites establecidos. Este enfoque de rejuvenecimiento se detalla en la Tabla 4.10.

Al reemplazar individuos con aptitud subóptima por otros recién generados, se fomenta sistemáticamente la heterogeneidad dentro de la población. Este método garantiza el mantenimiento de un nivel mínimo de calidad genética y reintroduce la variabilidad necesaria, facilitando una exploración continua y mitigando el riesgo de convergencias prematuras.

Tabla 4.10. Identificadores para el control de diversidad en algoritmos genéticos.

Identificador	Definición
`population`	Conjunto de individuos que forman la generación actual, evaluados para determinar la necesidad de intervenciones de diversidad genética.
`fitness`	Array que contiene las puntuaciones de aptitud de cada individuo en la población, utilizado para determinar qué individuos requieren regeneración.
`threshold`	Umbral de aptitud predefinido; los individuos cuya aptitud es inferior a este valor son candidatos para la renovación genética.
`parameter_ranges`	Diccionario que establece los rangos aceptables para cada parámetro genético, garantizando que las modificaciones permanezcan dentro de los límites viables.

4.8.7 Elitismo

El elitismo es una estrategia en los algoritmos genéticos que asegura la preservación de soluciones de alta calidad a lo largo de las generaciones. Este método protege a los individuos más aptos de ser perdidos debido a la selección aleatoria, el cruce, o la mutación. Se implementa ordenando a los individuos de la población por sus puntuaciones de aptitud mediante la función np.argsort(fitness), clasificándolos de menor a mayor. Los individuos más aptos, también conocidos como élites, son identificados y reservados para su inclusión en futuras generaciones.

El número de élites a conservar se determina mediante el parámetro num_elites. Estos individuos se seleccionan directamente del array ordenado, garantizando que las soluciones óptimas continúen influyendo en la evolución de la población. Esta práctica de extraer y mantener élites subraya la importancia de retener características genéticas superiores, minimizando el riesgo de regresión en la calidad de las soluciones a lo largo de sucesivas generaciones.

La Tabla 4.11 define los identificadores clave utilizados en la implementación del elitismo:

Tabla 4.11. Identificadores y sus definiciones en la función de elitismo.

Identificador	**Definición**
`population`	El conjunto total de individuos presentes en la generación actual.
`fitness`	Una lista que refleja las puntuaciones de aptitud para cada individuo dentro de la población.
`num_elites`	La cantidad de los mejores individuos seleccionados para preservación.
`elites_idx`	Índices dentro de la población que identifican a los individuos élites basados en su alta aptitud.
`elites`	Los individuos seleccionados que exhiben un rendimiento superior, reservados para futuras generaciones.

4.9 Identificación de temas por estado

En esta etapa, se obtuvieron los valores de los hiperparámetros α, β y Kutilizando un algoritmo genético. Este proceso se realizó para cada uno de los archivos correspondientes a cada entidad federativa, es decir, 32 archivos por cada año. En total, se procesaron 576 archivos para el período comprendido entre 2003 y 2020. Este análisis permitirá comprender las inquietudes de los ciudadanos en cada entidad geográfica y a lo largo de los años.

4.10 Interpretación de resultados

En el ámbito del análisis de texto mediante LDA, enfrentamos el reto de asignar títulos precisos y formular descripciones adecuadas para los temas identificados. Este desafío radica en la necesidad de interpretar y sintetizar de manera coherente y concisa las complejidades temáticas y los patrones extraídos a través del análisis LDA. Para abordar este problema, el estudio citado en (Rijcken et al., 2023) utiliza el modelo Transformer de Pre-entrenamiento Generativo (GPT) con el fin de mejorar la interpretación de los temas identificados.

Para mejorar la interpretación y descripción de los tópicos, nuestra metodología se basa en la utilización de palabras clave y sus respectivas probabilidades, lo que permite una comprensión más profunda de cada tema. La generación automática de títulos y descripciones se realiza mediante modelos GPT y se complementa con una fase de validación humana para garantizar la precisión y relevancia de los resultados.

En este proceso, las nubes de palabras desempeñan un papel fundamental, proporcionando una representación visual intuitiva que resalta los términos más significativos de cada tópico. Este enfoque no solo facilita la creación de títulos y descripciones más precisos, sino que también mejora la coherencia, claridad y estructura del texto, asegurando una conexión más profunda con el contenido analizado.

A continuación, se presenta el prompt utilizado en GPT-3.5-turbo para la generación automática de títulos y descripciones:

> *Eres un experto en análisis de textos y minería de datos. Se te proporcionará un conjunto de palabras clave extraídas mediante el modelo LDA, junto con sus ponderaciones. Tu tarea es interpretar el significado del tópico basándote en los términos más relevantes y sus relaciones.*
>
> *Para cada conjunto de palabras clave, debes analizar las palabras más frecuentes y su posible contexto para determinar el tema principal del tópico. Luego, genera un título conciso que resuma la idea central y redacta una breve descripción que explique el tema de manera clara y estructurada. Finalmente, clasifica el tópico dentro de una de las siguientes categorías: Adquisiciones, Comercial, Educación, Energía, Finanzas Públicas, Medio Ambiente, Necesidades Individuales, Salud, Seguridad o Servidores Públicos.*

A continuación,presentamos un ejemplo ilustrado en la Figura 4.6, La nube de palabras está relacionado con la pesca y regulaciones gubernamentales en México. Palabras destacadas como pesca, oficina, guía, veda, y octubre sugieren que podría tratarse de información sobre periodos regulados de pesca, procedimientos administrativos y pautas estacionales o temporales. También se mencionan localidades específicas como Campeche y Veracruz, lo que indica un enfoque regional en el tema. Otras palabras como sustentable y reproductivo aluden a la sostenibilidad y aspectos biológicos vinculados a la pesca. Para una interpretación más detallada, se utilizan palabras clave y su interpretación, que son procesadas a través del modelo GPT-3.5-turbo, como se muestra en la Tabla 4.12.

Figura 4.6. Ejemplo de tema: pesca sostenible y regulación en Campeche y Veracruz.

Tabla 4.12. Ejemplo de tema: título y descripción generados por GPT-3.5-turbo.

Título	**Palabras clave y Probabilidad**	**Descripción**
Pesca Sostenible y Regulación en Campeche y Veracruz	(pesca, 0.2600), (guía, 0.0330), (oficina, 0.0250), (pesquera, 0.0140), (virus sar, 0.0130), (octubre, 0.0110), (septiembre, 0.0090), (campeche, 0.0070), (especie, 0.0060), (dictamen, 0.0060), (rechazo, 0.0060), (veda temporal, 0.0050), (mexicano, 0.0050), (pacífico, 0.0040), (sustentable, 0.0040)	Este tema aborda la regulación y el control de la pesca en Campeche, destacando la importancia de las vedas temporales, la identificación de especies y la implementación de estándares sustentables. Incluye información sobre dictámenes y medidas de protección, así como actividades de revisión en el océano Pacífico y zonas cercanas. También se mencionan aspectos de certificación y manejo de virus SAR, con el objetivo de equilibrar explotación y conservación marina.

4.11 Clasificación de tópicos

Una etapa clave de esta investigación fue la clasificación de los tópicos, utilizando las categorías propuestas por (Berliner et al., 2022). En la Tabla 4.13 se detallan estas categorías y sus descripciones.

Tabla 4.13. Descripción de categorías de solicitudes adaptada de Berliner et al., 2022.

Categoría	**Descripción**
Medio Ambiente	Incluye proyectos y actividades relacionadas con el manejo y conservación del entorno natural. Las solicitudes pueden abordar temas como la gestión del agua, administración de terrenos, protección forestal, regulación de actividades mineras y otras preocupaciones ambientales.
Salud	Abarca temas relacionados con la salud pública y la atención médica. Incluye solicitudes sobre la infraestructura de hospitales, programas de salud pública, atención médica y servicios de emergencia.
Educación	Se refiere a las iniciativas y programas relacionados con el sistema educativo. Incluye solicitudes sobre la construcción y mantenimiento de escuelas, programas educativos, becas y oportunidades de capacitación.
Adquisiciones	Incluye todas las actividades relacionadas con la adquisición de bienes y servicios por parte del gobierno. Las solicitudes pueden tratar sobre contratos, licitaciones públicas, compras y adquisiciones.
Energía	Aborda cuestiones relacionadas con la producción, distribución y regulación de recursos energéticos. Incluye proyectos de energía renovable, políticas energéticas y la gestión de la electricidad.
Servidores Públicos	Se centra en el personal del sector público y su gestión. Incluye solicitudes sobre contrataciones, pensiones, capacitaciones y recursos humanos.
Seguridad	Trata temas relacionados con la protección y el orden público. Las solicitudes pueden referirse a la actuación de la policía, gestión de detenidos, operaciones militares, atención a víctimas y lucha contra la delincuencia y el uso de armas.
Comercial	Abarca actividades y regulaciones relacionadas con el comercio y la industria. Incluye solicitudes sobre aprobaciones regulatorias, inspecciones, propiedad intelectual y licencias de negocios.
Finanzas Públicas	Trata sobre la gestión de los recursos económicos del gobierno. Incluye solicitudes relacionadas con presupuestos, gastos, impuestos y programas distributivos.
Necesidades Individuales	Se refiere a solicitudes específicas de ciudadanos para acceder a programas gubernamentales o realizar trámites personales. Incluye aplicaciones a programas de gobierno, obtención de documentación oficial y otros servicios personalizados.

4.12 Aplicación de embeddings para detección de similitud

En esta etapa, se generan sentence-embeddings para las descripciones de los temas, que son representaciones vectoriales de la interpretación de los temas utilizando el modelo de incrustación de palabras conocido como sentence-embeddings-BETO. El proceso involucra calcular la similitud coseno entre cada par de descripciones de temas, estableciendo un umbral de similitud mayor a 0.8, posteriormente los temas se agrupan en función de su similitud utilizando el algoritmo K-means, lo que permite organizar y segmentar los temas junto con su respectivo estado. Finalmente, basado en los títulos agrupados, se asigna un nombre que engloba los temas a cada grupo. La Figura 4.7 muestra un ejemplo de similitud de temas por estado.

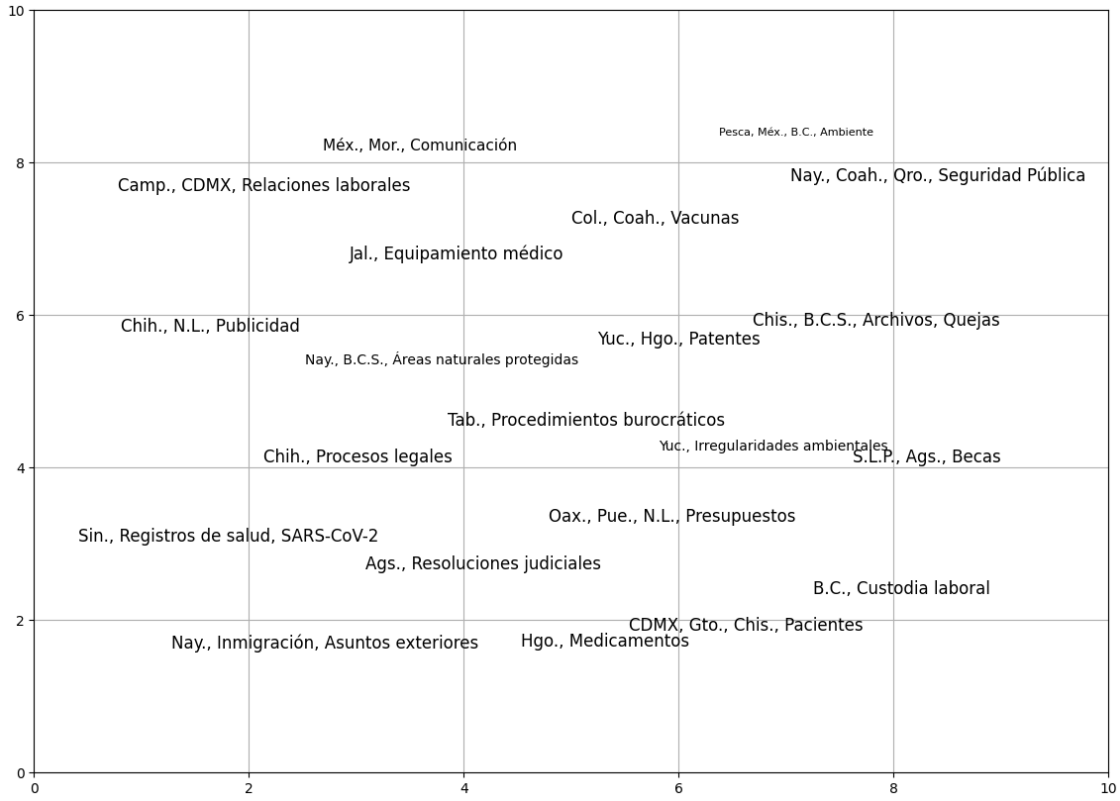

Figura 4.7. Ejemplo de similitud de temas.

4.13 Visualización de resultados

Finalmente, nuestro objetivo es presentar los hallazgos de este análisis utilizando herramientas visuales, como nubes de palabras, gráficos de barras y mapas geográficos. Todo esto puede integrarse en un dashboard interactivo que permita a cualquier persona interesada consultar los temas de interés para los ciudadanos.

5. Resultado y Discusión

5.1 Contexto solicitudes de información

Uno de los principales resultados de esta investigación fue el desarrollo de un dashboard llamado *Datainai*, disponible en https://datainai.streamlit.app/, una herramienta que participó en el Certamen de Innovación en Transparencia 2024. *Datainai* permite visualizar los tópicos de las solicitudes de información realizadas a INFOMEX y la Plataforma Nacional de Transparencia entre 2003 y 2020. En la Figura 5.1, se muestra una captura de pantalla de la aplicación, donde se destacan sus principales funcionalidades: un mapa coroplético para visualizar la distribución de solicitudes por estado, un mapa de dispersión para analizar las solicitudes según características como Sector, Respuesta y Estatus, y un mapa de calor que muestra la densidad de las solicitudes. Además, incluye un apartado con estadísticas generales de las solicitudes en ese periodo.

Figura 5.1. Interfaz del dashboard Datainai.

La Figura 5.2 muestra la cantidad de solicitudes generadas entre 2003 y 2020, sumando un total de 2,027,167. Desde 2003 hasta 2015, vemos un crecimiento constante en el número de solicitudes, con un notable aumento entre 2006 y 2007. Después, la tendencia sigue al alza, aunque de manera más gradual, hasta alcanzar un máximo en 2017. No obstante, en 2019 hay una caída en el número de solicitudes, seguida por una recuperación parcial en 2020.

La evolución en la cantidad de solicitudes de información mostrada en la Figura revela dinámicas significativas en la demanda de acceso a la información. El crecimiento constante observado desde 2003 hasta 2015 sugiere un aumento progresivo en la concienciación y el uso de mecanismos de transparencia por parte de los ciudadanos. El notable incremento entre 2006 y 2007 podría estar relacionado con cambios legislativos o políticas gubernamentales que incentivaron o facilitaron el acceso a la información (Barrios, 2017). Cabe destacar que, durante este periodo, se utilizaron diferentes plataformas para la gestión de solicitudes: entre 2003 y 2005, se empleó el Sistema de Solicitudes de Información (SISI), el cual fue reemplazado por INFOMEX en 2006, y posteriormente, en 2015, este sistema se transformó en la Plataforma Nacional de Transparencia (Martínez Díaz, Heras Gomez et al., 2011).

A partir de 2015, aunque la tendencia continúa al alza, el ritmo de crecimiento es más moderado, lo que podría indicar una saturación en el número de solicitudes o un acceso más eficiente a la información, reduciendo la necesidad de múltiples solicitudes. La caída en el número de solicitudes en 2019, seguida por una recuperación parcial en 2020, podría estar influenciada por factores externos, como crisis económicas o la suspensión de actividades, cuando el INAI decidió suspender sus labores, incluso de manera virtual, debido a la pandemia de COVID-19 (Rincón, 2020).

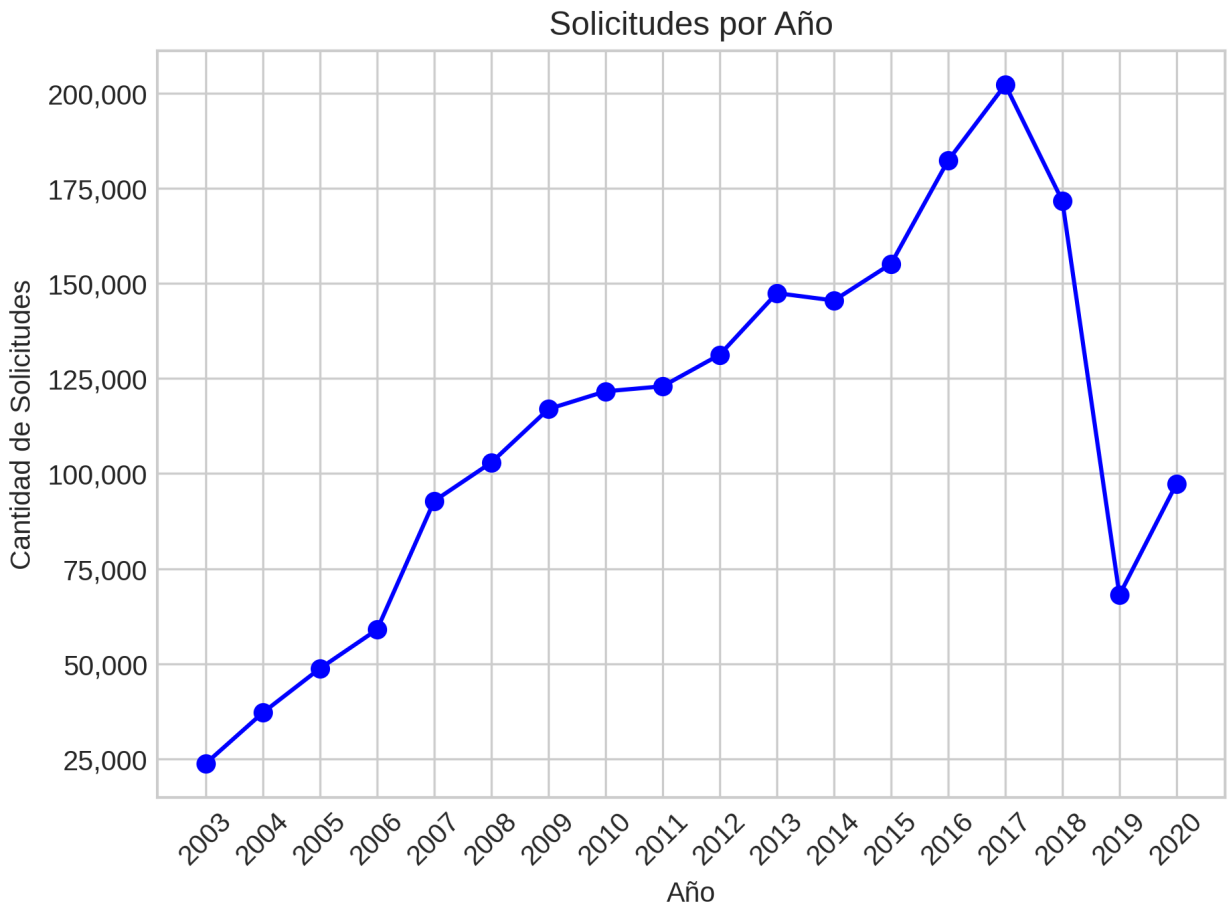

Figura 5.2. Número de solicitudes de información durante 2003–2020.

El análisis de los mapas, como se muestra en la Figura 5.3, revela un patrón general en la distribución geográfica de las solicitudes a lo largo de los años. Predominantemente, los estados de **Jalisco**, **Ciudad de México** y **Estado de México** concentran el mayor número de solicitudes. Esta concentración puede asociarse con su alta densidad poblacional, urbanización avanzada y la centralización de funciones administrativas y económicas, factores que tienden a intensificar la interacción ciudadana con las instituciones públicas y, por tanto, la demanda de información (Mayer-Foulkes, 2018; Padilla et al., 2023).

En contraste, entidades como **Chihuahua**, **Sonora**, **Tamaulipas**, **Quintana Roo** y **Veracruz** muestran consistentemente un volumen medio o bajo de solicitudes. Este comportamiento podría deberse a factores como la dispersión territorial, desigualdades en el acceso digital y diferencias en la apropiación social del derecho a la información.

Finalmente, los estados con el menor número de solicitudes, como **Baja California Sur**, **Durango**, **Zacatecas**, **Guerrero**, **Campeche**, **Nayarit** y **San Luis Potosí**, podrían enfrentar retos estructurales adicionales: menor infraestructura tecnológica, baja conectividad o reducidos niveles de alfabetización digital, lo que limita tanto el acceso como la demanda de información pública (Neri, 2022).

No obstante, es importante destacar ciertas *variaciones temporales* que emergen a lo largo del período analizado. Estados como **Baja California**, **Baja California Sur**, **Yucatán** y **Nuevo León** presentan fluctuaciones notables en distintos años, lo cual sugiere que factores contextuales como cambios institucionales, campañas locales de promoción del derecho a saber, o eventos coyunturales pueden incidir significativamente en la activación de solicitudes de información.

Figura 5.3. Mapas de solicitudes de información (2003–2020).

El mapa de la Figura 5.4 presenta una visualización coroplética de la distribución acumulada de solicitudes de información pública en México entre los años 2003 y 2020. A diferencia del análisis año por año, este mapa sintetiza la intensidad total de solicitudes por entidad federativa, lo que permite identificar de forma más clara patrones consolidados de uso del derecho de acceso a la información. Las tonalidades más oscuras reflejan una mayor concentración de solicitudes a lo largo del tiempo, destacando particularmente a la **Ciudad de México**, **Jalisco** y **Estado de México** como focos persistentes de demanda.

Por otro lado, esta representación también permite observar con mayor nitidez la persistencia de niveles bajos en entidades como **Campeche**, **Baja California Sur** y **Zacatecas**, las cuales se mantienen en los rangos inferiores del espectro. Asimismo, el mapa pone de relieve casos intermedios interesantes, como **Nuevo León**, **Yucatán** y **Chihuahua**, que si bien no lideran en términos absolutos, muestran un nivel de participación sostenido, posiblemente asociado con procesos locales de institucionalización del derecho o mejoras en el acceso digital. Este tipo de representación espacial acumulada ofrece una perspectiva integral que complementa el análisis de la evolución temporal, al evidenciar las trayectorias consolidadas de uso y rezago en el ejercicio del derecho a saber.

Figura 5.4. Distribución geográfica de solicitudes de información (2003–2020).

La Figura 5.5 muestra la distribución de las solicitudes por sector, elaborada a partir de los datos disponibles en la Plataforma Nacional de Transparencia (PNT). Entre los sectores que concentran el mayor número de solicitudes se encuentran: Aportaciones a Seguridad Social, Educación Pública,

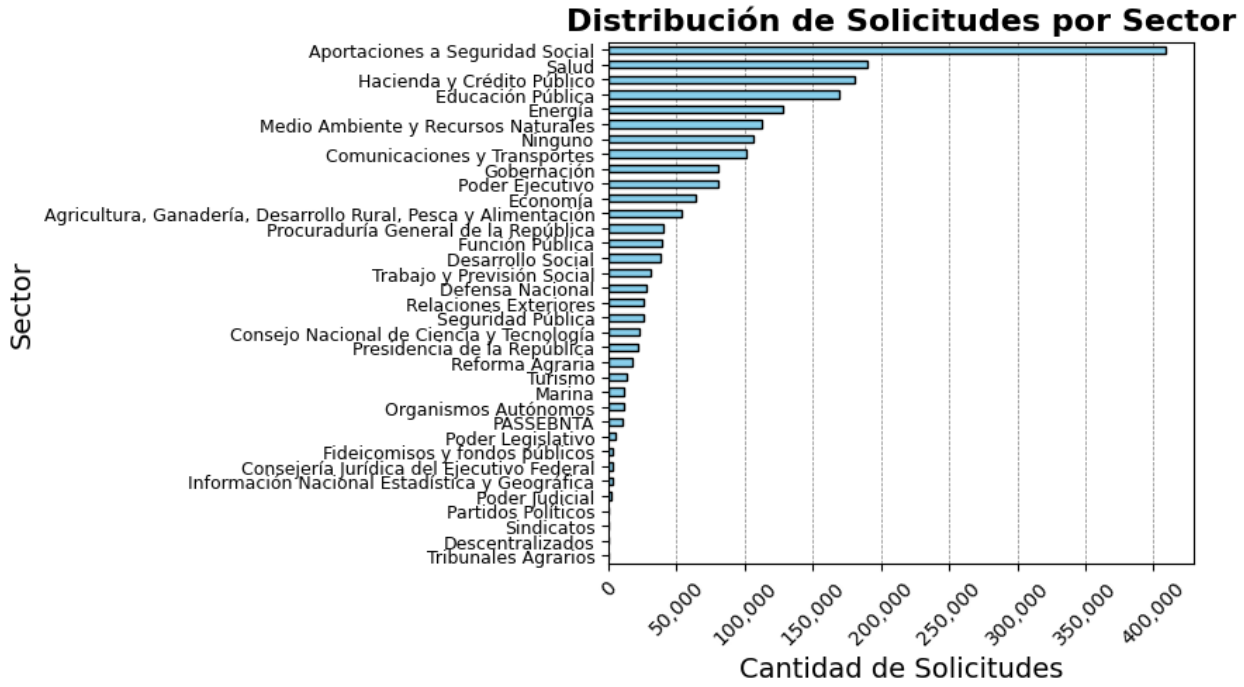

Figura 5.5. Solicitudes por sector (2003–2020).

En lo que respecta al medio empleado para realizar las solicitudes de información, la mayoría de ellas, específicamente 1,941,757, se efectúan de manera electrónica a través de la Plataforma Nacional de Transparencia. Por otro lado, un total de 85,410 solicitudes se gestionan manualmente, lo cual generalmente implica la asistencia presencial a la unidad de enlace correspondiente ver Figura 5.6.

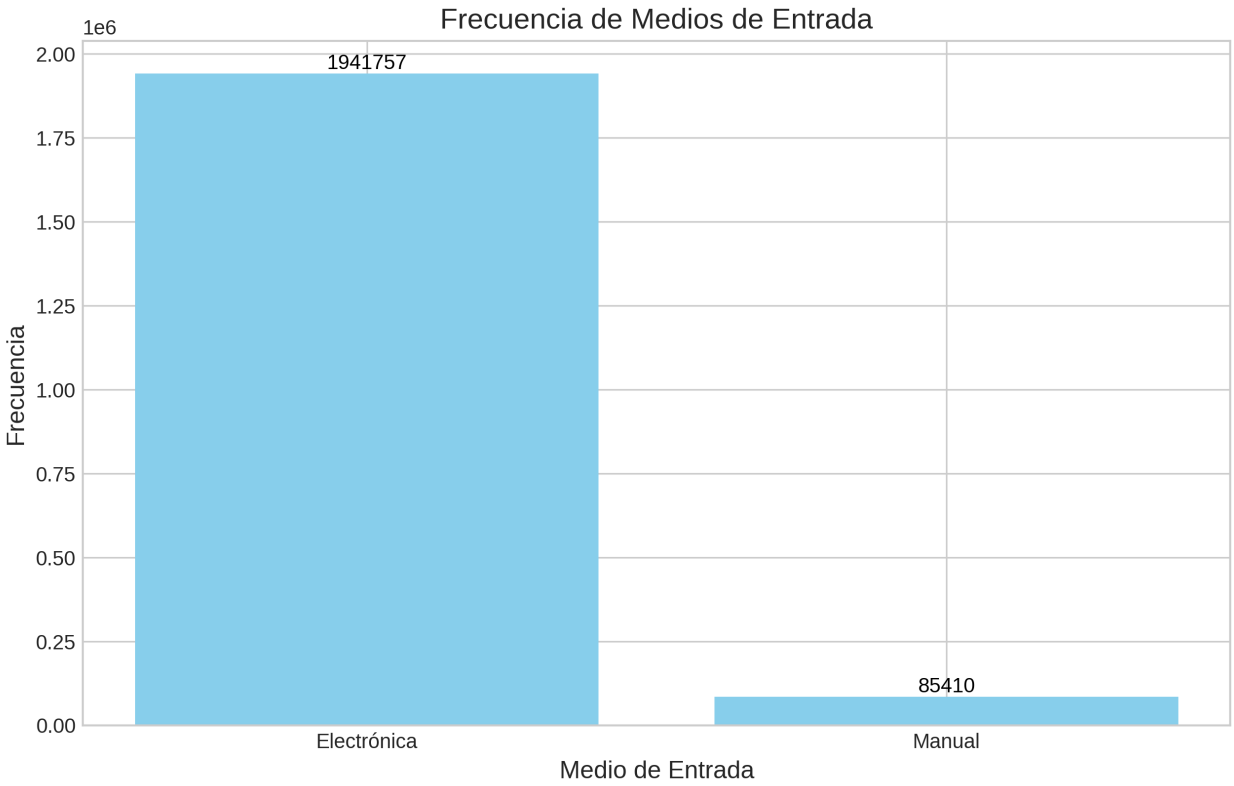

Figura 5.6. Solicitudes por medio de entrada (2003–2020).

En relación con el tipo de solicitud presentada ver Figura 5.7, un total de 1,611,358 individuos realizaron solicitudes de información pública a diversas entidades del ámbito federal. Por otro lado, 415,809 personas efectuaron solicitudes de acceso a información relacionada con sus datos personales. Esto último implica el ejercicio de los derechos ARCO, término que engloba los derechos de Acceso, Rectificación, Cancelación, Oposición y Portabilidad de datos. Dichos derechos permiten al titular de los datos personales solicitar ante el ente responsable o Sujeto Obligado, la gestión adecuada de su información personal conforme a los preceptos legales establecidos.

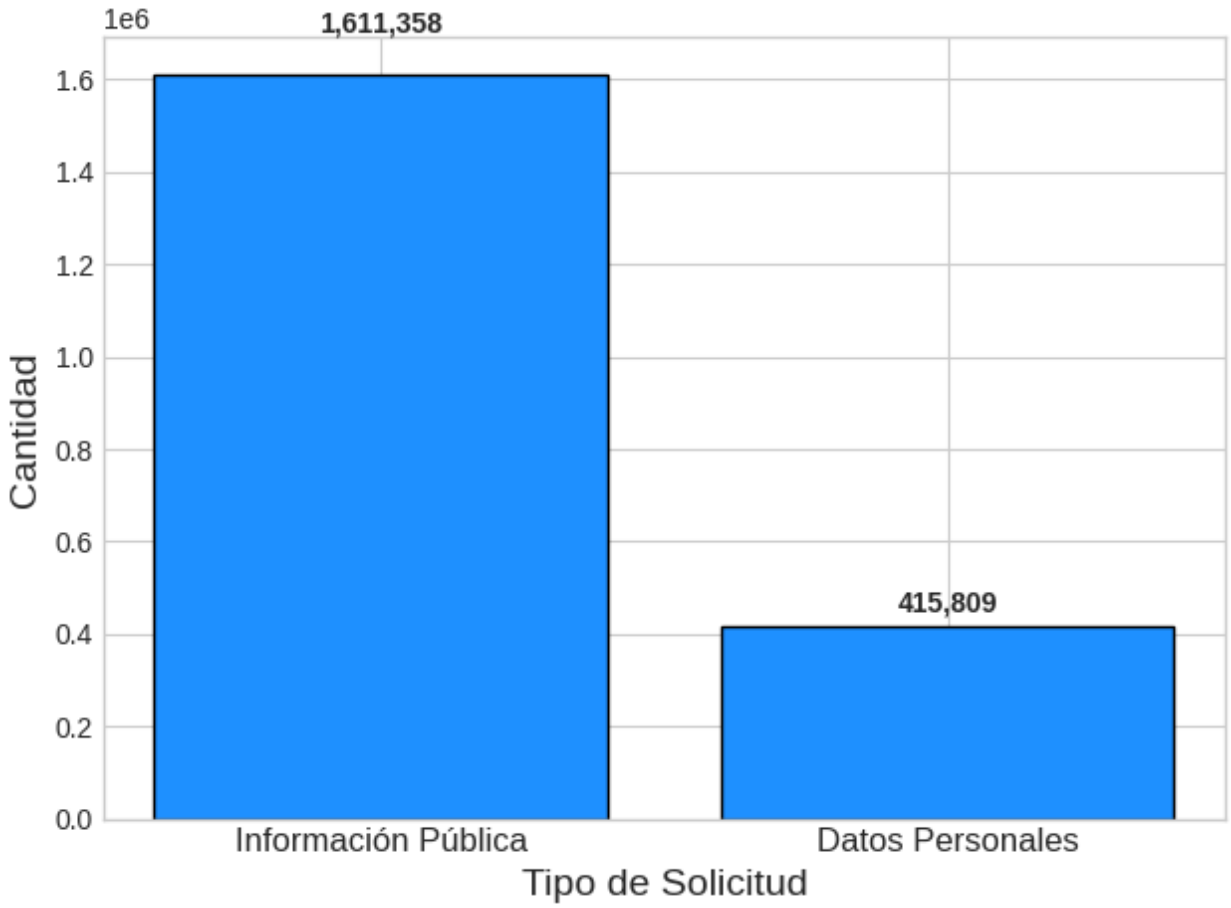

Figura 5.7. Solicitudes por tipo (2003–2020).

La distribución de los métodos de entrega de información, ilustrada en la Figura 5.8, evidencia la preferencia por las soluciones digitales, con Entrega por Internet en el INFOMEX y Entrega por Internet en la PNT predominando claramente con 881,602 y 523,880 solicitudes, respectivamente. Esta tendencia subraya la eficiencia y accesibilidad que las plataformas digitales ofrecen. Aunque en una escala menor, la demanda de copias certificadas y simples aún es significativa, lo que señala la persistente necesidad de documentación física en determinadas circunstancias. Los métodos menos frecuentes, como la consulta directa, la entrega en disco o CD, y la comunicación verbal, indican un interés decreciente por formatos más tradicionales o menos automatizados.

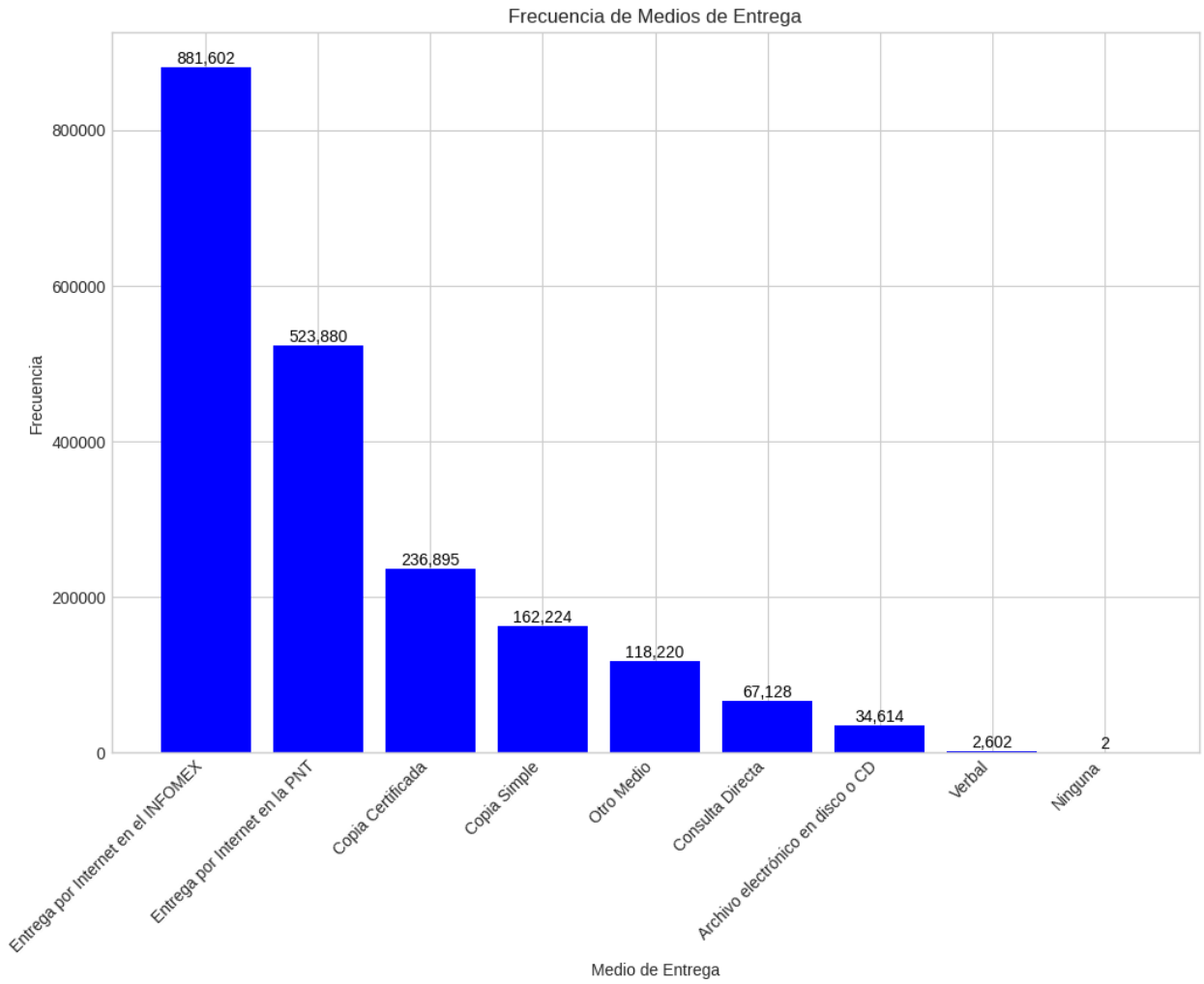

Figura 5.8. Solicitudes por medio de entrega (2003–2020).

5.2 Resultados de la ley de Zipf

La Figura 5.9 muestra la verificación del cumplimiento de la Ley de Zipf en la distribución del vocabulario extraído de las solicitudes de información pública en el estado de Jalisco durante el año 2020. La curva representa la frecuencia de aparición de las palabras en función de su rango, siguiendo el principio característico de dicha ley. La línea azul, etiquetada como Antes del Punto Crítico, destaca cómo las palabras más comunes dominan el vocabulario. Estas palabras, aunque frecuentes, suelen ser genéricas y no aportan un valor significativo en el contexto de las solicitudes de información pública. Por esta razón, se ha decidido eliminarlas durante el proceso de optimización del vocabulario.

El punto crítico, señalado en el gráfico con una flecha y anotado como S=1.00, marca el umbral donde se realiza la selección. Después de este punto, la línea roja Después del Punto Crítico representa las palabras que se conservarán en el vocabulario. Estas palabras, aunque menos frecuentes, son más específicas y relevantes, y por lo tanto, se consideran esenciales para un análisis más profundo y preciso.

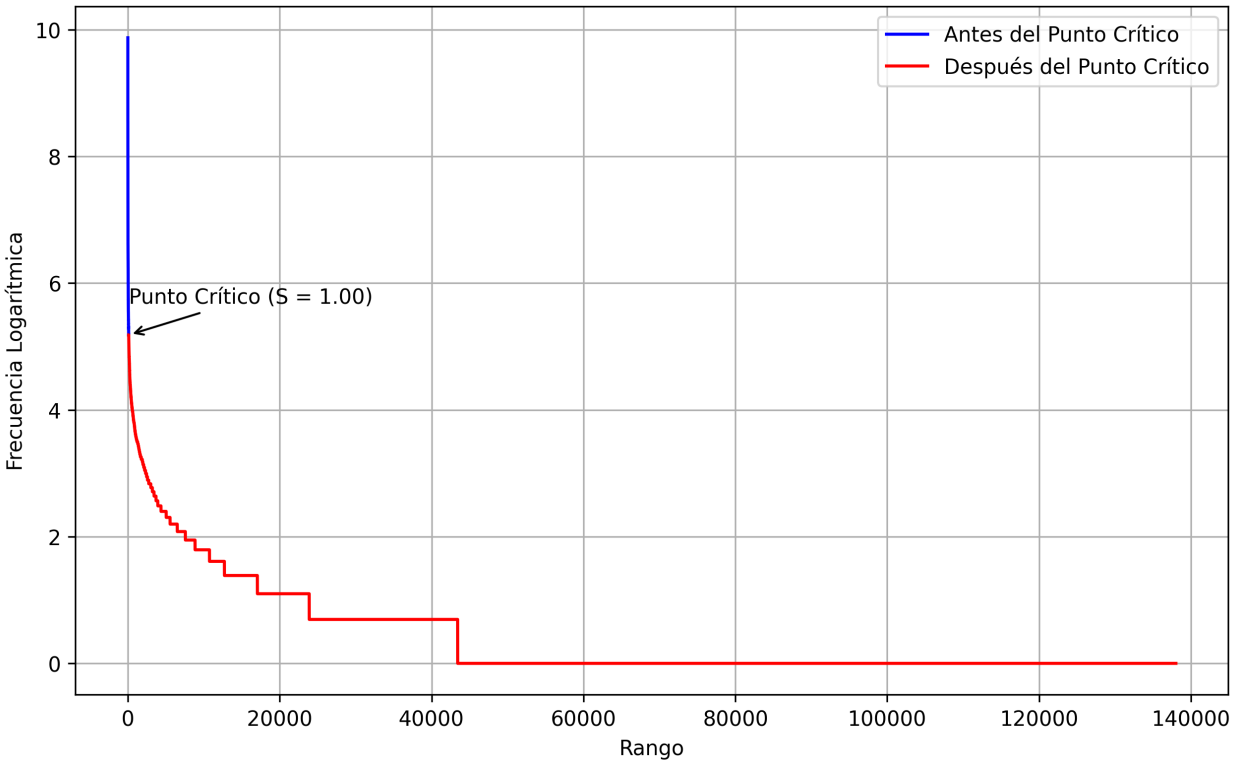

Figura 5.9. Resultado de la ley de Zipf.

La aplicación de la Ley de Zipf en el análisis de las solicitudes de acceso a la información pública ha demostrado ser una estrategia eficaz para optimizar el vocabulario y mejorar la precisión del análisis textual. Al eliminar las palabras más comunes y genéricas, que dominan el corpus sin aportar un valor analítico significativo, se ha logrado un enfoque más centrado en términos específicos y relevantes, lo que reduce el ruido en los datos y permite un análisis más profundo y preciso. Estos resultados coinciden con estudios previos que han validado la utilidad de la Ley de Zipf para mejorar la calidad y relevancia del análisis en grandes corpus textuales (Piantadosi, 2014).

Además, la Ley de Zipf ha sido útil para identificar y definir listas de stop words en el contexto de las solicitudes de información, eliminando términos legales recurrentes y expresiones de cortesía que, aunque comunes, no aportan información relevante para la clasificación temática. También ha facilitado la detección de errores ortográficos y palabras mal escritas, mejorando la limpieza y precisión del análisis al eliminar elementos que podrían distorsionar los resultados.

La Figura 5.10 muestra una comparación de las 25 palabras más frecuentes antes y después de aplicar el análisis del punto de rodilla para optimizar el vocabulario. En el gráfico de la izquierda, que muestra la distribución antes de la optimización, predominan palabras genéricas y funcionales como “del”, “que”, “por”, y “se”, así como términos relacionados con la estructura administrativa y el proceso de solicitud, como “solicito”, “solicitud”, y “número”. Estas palabras, aunque comunes, pueden no ofrecer un valor analítico profundo, ya que son frecuentemente utilizadas en un amplio rango de contextos sin una referencia específica a los temas más relevantes de las solicitudes de información pública. Por otro lado, el gráfico de la derecha, que presenta los resultados después de la optimización, destaca un vocabulario más centrado en temas clave y específicos de las solicitudes. Palabras como “servicios”, “caso”, “registro”, “sistema”, y “proyecto” sugieren que el proceso de optimización ha eliminado las palabras menos relevantes, permitiendo un enfoque más claro en los temas sustantivos de interés público.

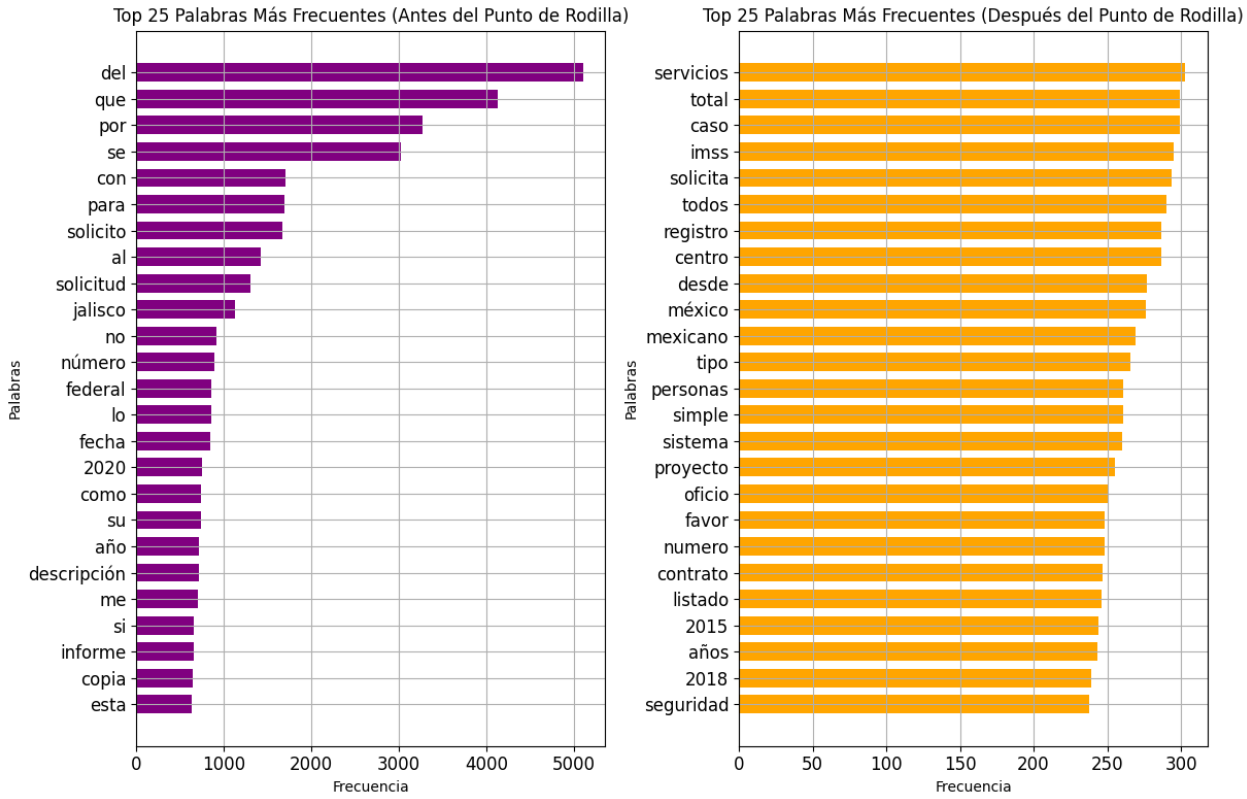

Figura 5.10. Frecuencia de palabras antes y después del punto de rodilla.

5.3 Evolución del algoritmo genético

La Figura 5.11 ilustra un ejemplo de la evolución de la aptitud dentro de un algoritmo genético a lo largo de 50 generaciones, aplicado a un conjunto de 2,580 solicitudes de información del estado de Puebla. Se nota una tendencia ascendente en las primeras etapas, lo que sugiere que el algoritmo fue capaz de identificar rápidamente candidatos prometedores y mejorar significativamente la calidad de las soluciones en comparación con las generaciones iniciales. Después de este aumento inicial, la curva de aptitud se estabiliza, lo cual es típico en los algoritmos genéticos, ya que la población tiende a converger hacia una solución óptima local o global.

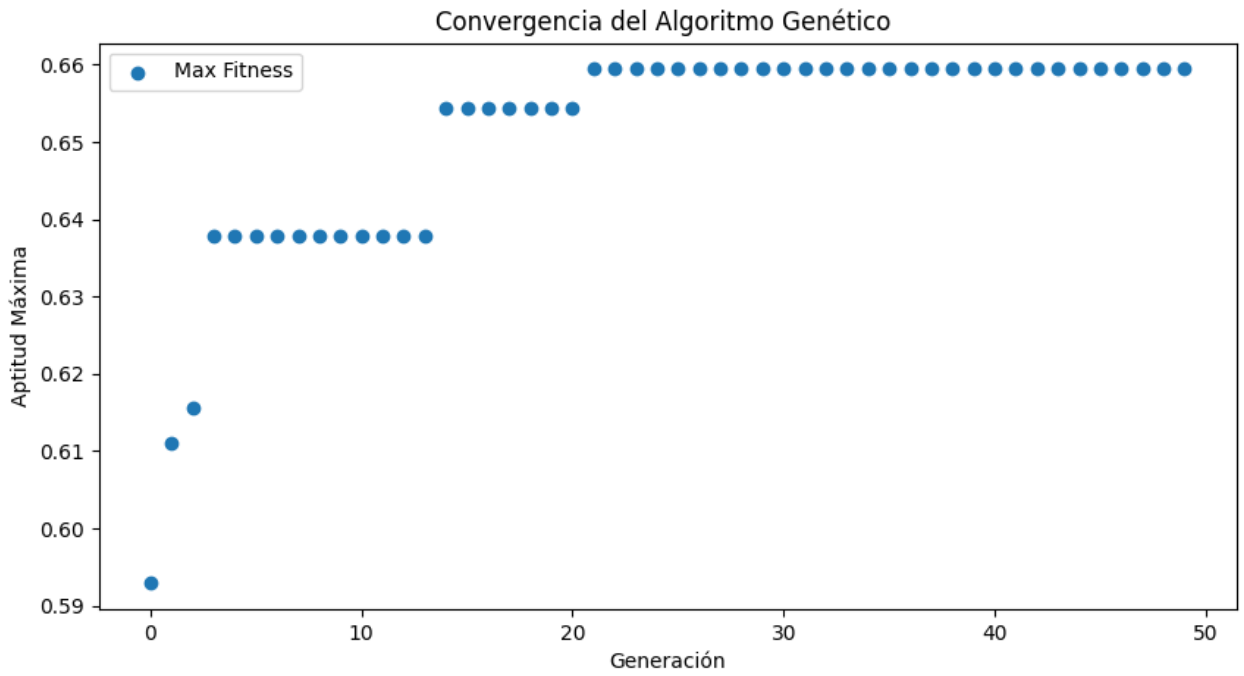

Figura 5.11. Evolución del rendimiento a lo largo de las generaciones.

El mejor individuo que surgió de este proceso mostró una configuración de parámetros con un α de 0.18 y un β de 0.81, manejando 8 temas diferentes en su modelo. Este conjunto de parámetros alcanzó un valor de coherencia de 0.73, lo cual es un indicador sólido de que la solución encontrada es significativa y bien estructurada, según las métricas del modelo LDA utilizadas para medir la coherencia temática. Este valor sugiere que los temas generados son claramente diferenciables entre sí, capturando conjuntos distintos de información dentro del corpus analizado, aunque no necesariamente mutuamente excluyentes, ya que puede existir cierto solapamiento natural entre algunos de ellos.

La Tabla 5.1 presenta los valores de los hiperparámetros α, β y K, junto con sus respectivos valores de coherencia para cada entidad federativa durante el año 2020. Un análisis detallado de esta Tabla revela una tendencia interesante: si bien en algunos casos se observa que un mayor número de solicitudes corresponde a un incremento en el número de temas, esta relación no se mantiene uniforme en todas las instancias. Este hallazgo sugiere una variabilidad en la distribución de temas que no está directamente correlacionada con el volumen de solicitudes en todas las entidades. Para más detalles de los años 2003 a 2019, consulte las tablas en el apéndice.

Tabla 5.1. Resultados de hiperparámetros por entidad en 2020.

Entidad	No. Solicitudes	α	β	K	Coherencia
Ags.	533	0.28	0.49	5	0.62
B.C.	2,540	0.68	0.22	9	0.84
B.C.S.	483	0.39	0.43	4	0.70
Camp.	205	0.90	0.39	4	0.74
Chis.	370	0.71	0.66	5	0.61
Chih.	1354	0.85	0.87	4	0.52
CDMX	324,948	0.26	0.79	56	0.82
Coah.	651	0.56	0.45	10	0.54
Col.	234	0.67	0.47	6	0.59
Dgo.	346	0.85	0.60	3	0.60
Mex.	9,740	0.28	0.75	9	0.66
Gto.	916	0.90	0.17	9	0.50
Gro.	351	0.86	0.54	7	0.65
Hgo.	859	0.77	0.46	4	0.51
Jal.	4,109	0.92	0.60	5	0.46
Mich.	386	0.20	0.19	3	0.64
Mor.	904	0.99	0.14	5	0.52
Nay.	633	0.30	0.61	3	0.51
N.L.	1,111	0.91	0.95	7	0.45
Oax.	554	0.84	0.35	6	0.49
Pue.	2,580	0.18	0.81	8	0.73
Qro.	1,503	0.76	0.84	9	0.54
Q. Roo	847	0.18	0.76	4	0.60
S.L.P.	410	0.53	0.93	3	0.64
Sin.	1,590	0.51	0.60	7	0.55
Son.	1,139	0.91	0.27	6	0.63
Tab.	2,150	0.23	0.48	3	0.55
Tamps.	1,003	0.72	0.84	7	0.58
Tlax.	480	0.05	0.81	4	0.68
Ver.	1,767	0.85	0.10	7	0.48
Yuc.	1,339	0.72	0.94	8	0.58
Zac.	182	0.29	0.07	5	0.51

Fuente: Elaboracion propia con datos del INAI (2023).

Al comparar nuestra metodología, que utiliza algoritmos genéticos para optimizar los parámetros α, β y K en función de la coherencia, con el enfoque de (Berliner et al., 2018), que emplea Gibbs sampling para estimar un modelo LDA de 20 temas sobre un corpus extenso de solicitudes de información pública en México, se observan diferencias significativas tanto en el proceso de modelado como en la interpretación de los resultados. Mientras que nuestra metodología se enfoca en la optimización de α, β y K para maximizar la coherencia, asegurando que los temas resultantes sean más coherentes y semánticamente significativos, el enfoque basado en Gibbs sampling identifica temas a partir de la probabilidad posterior, complementado con un análisis de palabras frecuentes y exclusivas (FREX) para su interpretación. Además, en la clasificación de documentos, mientras que el enfoque de Gibbs sampling asigna documentos según el tema con mayor probabilidad de asociación, nuestro método optimiza la clasificación para que esté mejor alineada con la coherencia semántica. Finalmente, aunque ambos enfoques abordan la dinámica temporal de los temas, nuestra metodología ofrece una representación más detallada al considerar cambios en la representatividad de los temas a lo largo del tiempo y por ubicación geográfica, ya que los parámetros α, β y K fueron optimizados para cada una de las entidades de la República Mexicana durante el período 2003-2020. Además, para la asignación de títulos, descripciones y categorías, utilizamos modelos de lenguaje como GPT-3.5 Turbo para la interpretación automática basada en probabilidades, lo que añade una capa adicional de precisión y relevancia en la categorización de los temas.

5.4 Resultados de tópicos

En esta sección se presentan los tópicos identificados a través de la metodología propuesta. Los resultados del modelado de tópicos se exponen de manera global, abarcando el período comprendido entre 2003 y 2020, y también se desglosan por zonas económicas del país.

5.4.1 Resultado modelado de tópicos global

La Figura 5.12 muestra la cantidad de tópicos identificados entre 2003 y 2020, obtenidos mediante nuestra metodología propuesta. Se observa una tendencia creciente en la cantidad de tópicos a partir de 2009, alcanzando un pico máximo en 2017. Posteriormente, hay una disminución hasta 2018, seguida de una ligera recuperación en 2020.

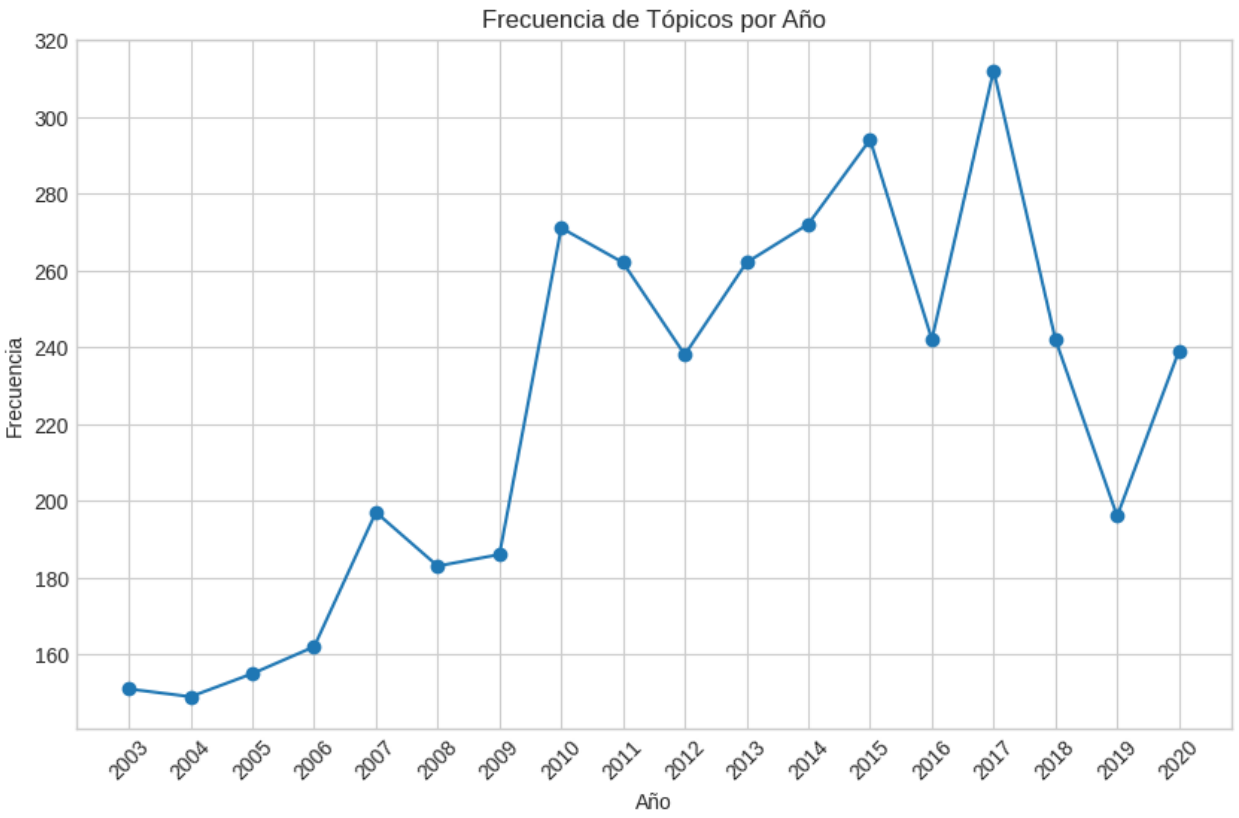

Figura 5.12. Número de tópicos durante el periodo 2003–2020.

La Figura 5.13 ilustra las diez categorías temáticas derivadas del análisis de tópicos, las cuales fueron clasificadas manualmente a partir de la interpretación de los tópicos generados por el modelo LDA. Esta clasificación no se basó en la lectura individual de cada solicitud, sino en el análisis cualitativo de las palabras clave más representativas y sus probabilidades asociadas en cada tópico. Cabe destacar que las categorías utilizadas no fueron arbitrarias: se tomaron como referencia las propuestas por expertos en el análisis de solicitudes de información pública, particularmente el esquema temático desarrollado por (Berliner et al., 2022).

Dentro de las categorías analizadas, **Servidores Públicos** resalta como la de mayor volumen general, registrando los **picos más altos en los años 2010, 2015 y 2017**, con una disminución posterior especialmente visible a partir de 2018. Esta categoría domina en varias etapas del periodo, lo que sugiere una alta recurrencia de solicitudes o eventos relacionados con esta temática. En contraste, las categorías de **Medio Ambiente** y **Seguridad** presentan **tendencias ascendentes moderadas** a lo largo del tiempo, con momentos de aumento notables en **2010, 2014 y 2017**, lo cual podría reflejar una creciente atención hacia estas áreas en la agenda pública.

Por otro lado, **Comercial** y **Salud** muestran **comportamientos más irregulares y de menor volumen**, caracterizados por fluctuaciones puntuales en años como **2004, 2010 y 2015**, lo que podría estar asociado a contextos o necesidades específicas surgidas en esos periodos.

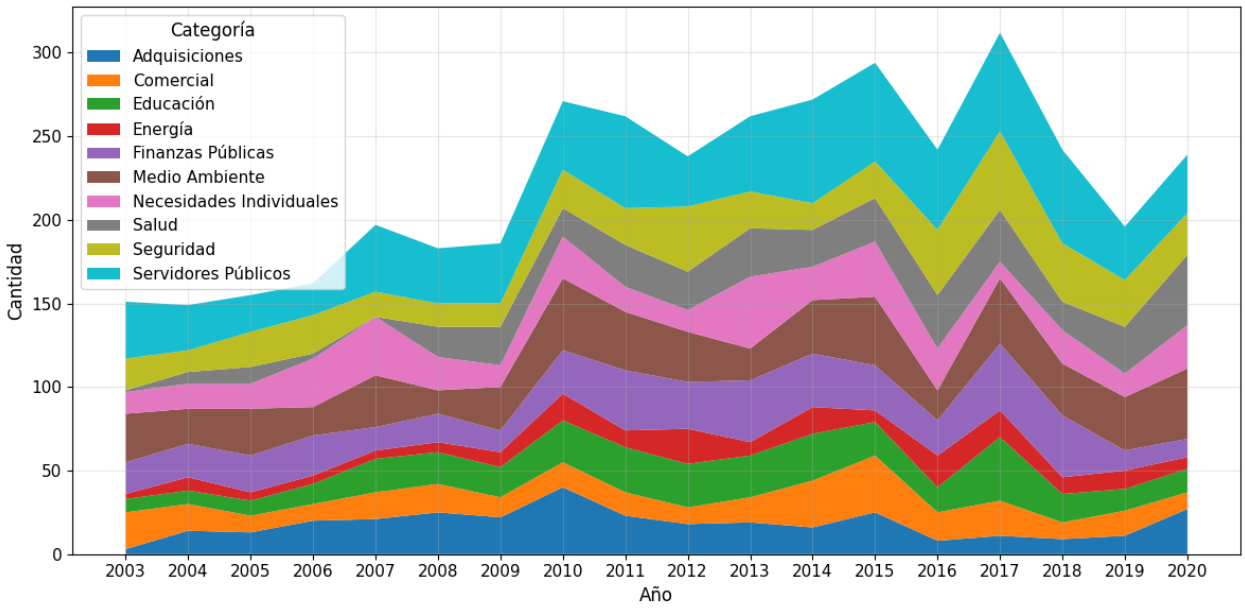

Figura 5.13. Tendencia de temas en solicitudes de información (2003–2020).

5.4.2 Zona económica noroeste

La Tabla 5.2 muestra que para la Zona Económica Noroeste se identificaron 572 tópicos en total, siendo Sinaloa el estado con el mayor número de tópicos (109), seguido de Baja California (101). Por otro lado, Durango tiene el menor número de tópicos (84).

Tabla 5.2. Número de tópicos por estado en la zona económica noroeste durante el periodo 2003-2020.

Estado	**Número de Tópicos**
Baja California	101
Baja California Sur	91
Chihuahua	94
Durango	84
Sinaloa	109
Sonora	92

En la Figura 5.14, se aprecia que la categoría de **Servidores Públicos** presenta **picos significativos en los años 2007, 2011 y 2017**, seguidos de una disminución progresiva en los años posteriores. La categoría de **Medio Ambiente** evidencia una **tendencia creciente**, alcanzando **sus niveles más altos en 2017 y 2018**, lo que sugiere un aumento en la atención hacia esta temática en la región.

Por otro lado, la categoría de **Educación** muestra **picos destacados en 2010 y 2017**, lo que podría estar vinculado a políticas públicas o eventos relevantes en esos periodos. Finalmente, la categoría de **Seguridad** registra un **aumento notable en 2017**, aunque en términos generales mantiene una frecuencia relativamente baja en comparación con otras categorías.

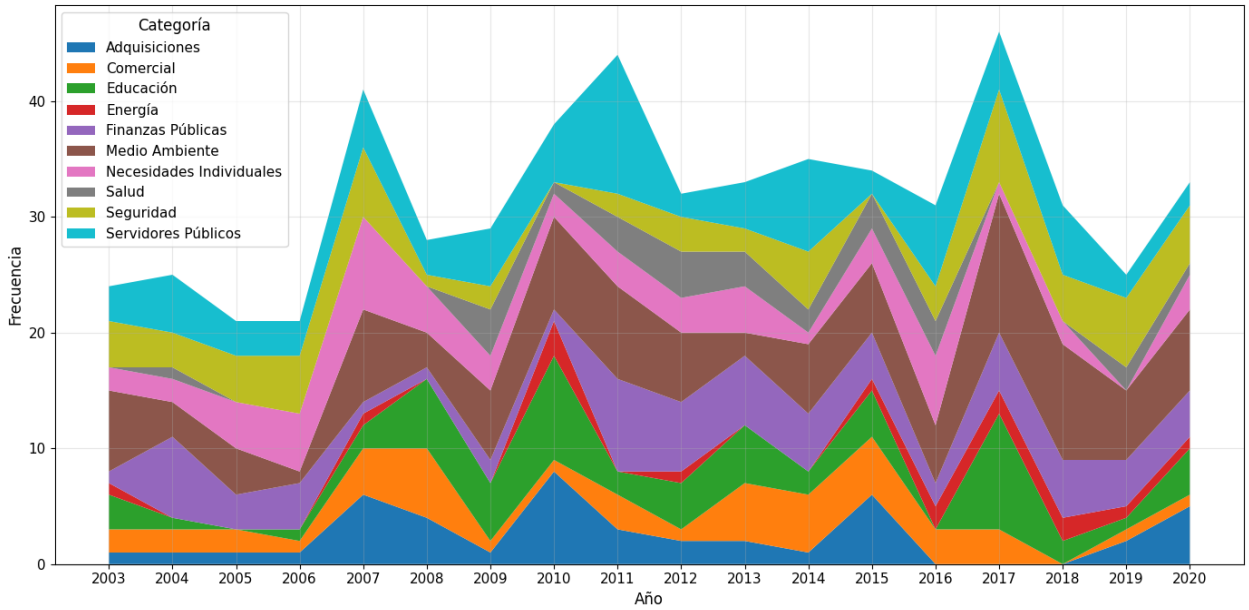

Figura 5.14. Tendencia de categorías durante el periodo 2003–2020 en la zona noroeste.

Ejemplos de tópicos representativos zona noroeste

La Tabla 5.3 muestra ejemplos de tópicos de la zona noroeste de México. En 2003, Baja California se enfoca en los procedimientos para concesiones y licencias en áreas de venta y radio. En 2018, Baja California Sur analiza la gestión de actividades extractivas y de conservación por parte de la SEMARNAT, con énfasis en la observación de tiburones ballena. Chihuahua, en 2010, aborda la comunicación celular y proyectos federales, resaltando la participación de individuos específicos.

Durango, en 2012, se centra en pensiones y beneficios sociales, destacando programas de apoyo para trabajadores y mujeres. Sinaloa, en 2007, investiga contratos en el sector federal y preventivo, con un enfoque en medidas preventivas y prestaciones laborales en Mazatlán. Sonora, en 2004, analiza inversiones y programas a nivel estatal y nacional, con atención especial a Guerrero y cambios durante la administración de Zedillo. Esta Tabla proporciona una visión general de las prioridades y preocupaciones en la zona noroeste de México, mostrando variaciones en los enfoques y temas entre diferentes estados y años. Para más detalles de los años 2003 a 2019, consulte las Tablas en el Apéndice A.

Tabla 5.3. Tópicos representativos zona noroeste

Año	Estado	Topic	Título	Probabilidades	Descripción	Categoría
2003	B.C.	3	Requisitos y Procedimientos para Concesiones y Licencias en Áreas de Venta y Radio	Área (0.0280), venta (0.0270), requisito (0.0230), estación (0.0230), concesión (0.0230), requisitos (0.0230), radio (0.0230), origen (0.0230), cubrir (0.0230), tramitar (0.0230), ...	se enfoca en los requisitos y procedimientos para obtener concesiones y licencias en áreas de venta y radio. Examina los requisitos específicos para tramitar licencias y concesiones, especialmente en estaciones de radio y áreas de venta, incluyendo condominios y propiedades comunes.	Adquisiciones
2018	B.C.S.	1	SEMARNAT y la Gestión de Actividades Extractivas y de Conservación	SEMARNAT (0.0340), actividades (0.0260), artículo (0.0250), número (0.0220), cargo (0.0170), entregue (0.0120), base (0.0120), cardones (0.0120), tiburón ballena (0.0110), observación nado (0.0110), ...	se centra en la SEMARNAT y su papel en la gestión de actividades extractivas y de conservación, especialmente relacionadas con el nado y observación del tiburón ballena en bahías.	Medio Ambiente
2010	Chih.	4	Comunicación Celular y Proyectos Federales	celular (0.0390), simple (0.0340), Sra. (0.0290), octubre (0.0280), meses (0.0240), mayo junio (0.0210), Alejandro Clemente (0.0200), Cristo (0.0150), federal (0.0140), Lourdes (0.0140), ...	Enfoca en la comunicación celular y proyectos federales. Celular, simple y Sra. sugieren un enfoque en la comunicación celular y la participación de individuos específicos. Octubre, mes y mayo junio indican un contexto temporal específico en relación con estos proyectos. Alejandro Clemente, cristo y federal resaltan la participación de personas específicas en proyectos a nivel federal.	Energía
2012	Dgo.	0	Pensiones y Beneficios Sociales	general (0.0530), trabajadores (0.0380), federales (0.0350), federal (0.0280), social (0.0270), anterior (0.0200), misma (0.0200), apartado (0.0180), funcionario (0.0160), cooperativas (0.0150), ...	Este tópico se centra en las pensiones y otros beneficios sociales, discutiendo las bases y condiciones para obtener mejoras conforme a diferentes situaciones laborales y sociales. Incluye menciones a recursos adicionales para trabajadores y mujeres, así como programas específicos que ofrecen apoyo adicional, como estancias infantiles y rehabilitación.	Necesidades Individuales
2007	Sin.	4	Investigaciones y Contratos en el Sector Federal y Preventivo	federal (0.0350), trabajadores (0.0170), preventiva (0.0160), Mazatlán (0.0150), investigaciones (0.0090), domicilio (0.0090), divide (0.0090), tarde (0.0090), gobierno (0.0070), contrato colectivo (0.0050), ...	Este tópico aborda investigaciones y contratos en el sector federal, con un enfoque en medidas preventivas y desarrollo de trabajadores. Examina la situación en Mazatlán, incluyendo aspectos domiciliarios y divisiones de trabajo. Menciona áreas como saneamientos sanitarios, expedientes de trabajo y prestaciones.	Adquisiciones
2004	Son.	0	Inversiones y Programas en Estados y Municipios	Estado (0.0770), Nacional (0.0320), programa (0.0250), inversiones (0.0180), año (0.0180), guerrero (0.0160), oportunidades (0.0150), específicamente (0.0120), cuatro (0.0120), también (0.0120), ...	Este tópico se centra en las inversiones y programas a nivel estatal y nacional, con especial atención a Guerrero y otras localidades. Se examinan las oportunidades específicas y los cambios durante la administración de Zedillo. Se discute la participación en programas populares y rurales, con mención a astilleros y desarrollo marítimo.	Finanzas públicas

5.4.3 Zona económica noreste

La Tabla 5.4 muestra la cantidad de tópicos de la Zona Económica Noreste. En total, se identificaron 312 tópicos, siendo Nuevo León el estado con el mayor número de tópicos (113), seguido de Coahuila con 104 y, finalmente, Tamaulipas con 95.

Tabla 5.4. Número de tópicos por estado en la zona económica noreste durante el periodo 2003–2020.

Estado	Número de Tópicos
Coahuila	104
Nuevo León	113
Tamaulipas	95

La Figura 5.15, destaca un **pico general en 2011**, impulsado principalmente por las categorías de **Educación**, **Comercial** y **Servidores Públicos**. Esta última categoría mantiene una alta frecuencia a lo largo del tiempo, con repuntes importantes en **2014**, **2015** y **2020**. La categoría de **Medio Ambiente** muestra una **tendencia creciente**, alcanzando su punto máximo en **2020**. Por su parte, **Seguridad** incrementa su presencia entre **2016 y 2019**, mientras que **Finanzas Públicas** y **Salud** se mantienen relativamente estables. Otras categorías, como **Adquisiciones** y **Energía**, presentan comportamientos más irregulares, con picos puntuales.

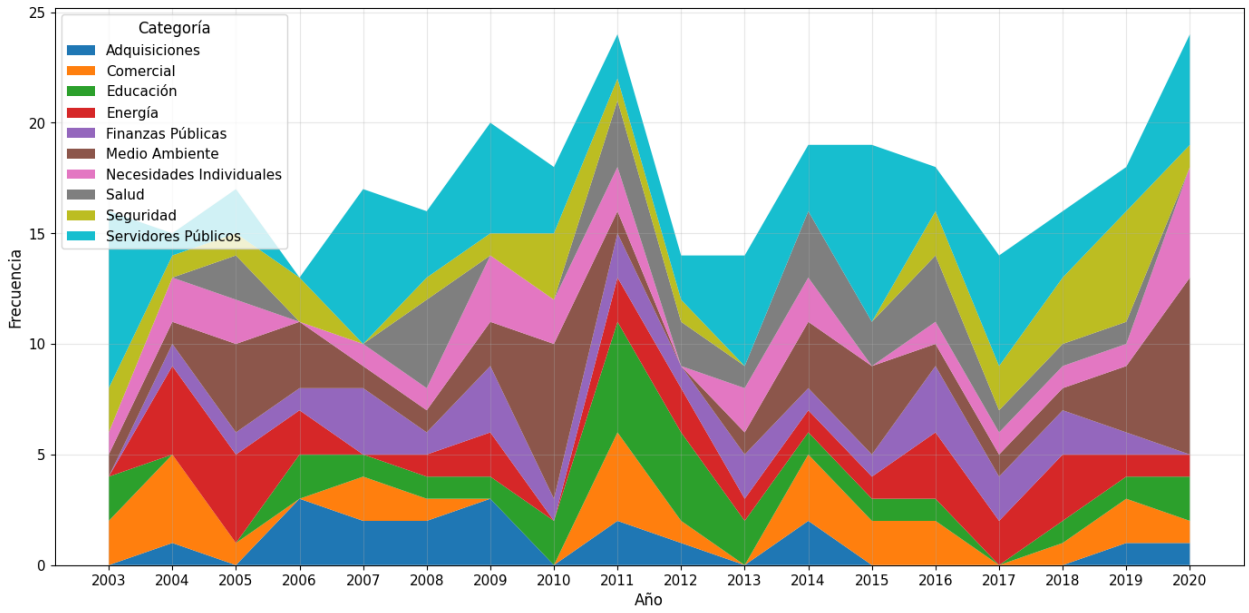

Figura 5.15. Tendencia de categorías durante el periodo 2003–2020 en la zona noreste.

Ejemplos de tópicos representativos zona noreste

La Tabla 5.5 presenta una síntesis de los tópicos representativos de la zona noreste de México en distintos años y estados, abarcando temas diversos como adquisiciones, finanzas públicas y salud. En el año 2007, en Coahuila, el tópico 2 se enfoca en las concesiones y recursos federales, resaltando las interacciones con el gobierno federal y la gestión de desastres naturales. Para el año 2019, en Nuevo León, el tópico 8 trata sobre fiscalización y estudios a nivel nacional, poniendo énfasis en la importancia de los transbordadores y la impugnación de decisiones. Finalmente, en el año 2018, en Tamaulipas, el tópico 0 aborda las políticas federales en salud y educación, destacando la regulación sanitaria y las colaboraciones bilaterales. Para más detalles de los años 2003 a 2019, consulte las Tablas en el Apéndice A.

Tabla 5.5. Tópicos representativos zona noreste

Año	Estado	Topic	Título	Probabilidades	Descripción	Categoría
2007	Coah.	2	Concesiones y Recursos Federales en Coahuila	Coahuila (0.0520), cuáles (0.0400), federales (0.0360), cuánta (0.0320), concesiones (0.0250), principales (0.0170), estimulo (0.0140), quisiera (0.0110), contratos monto (0.0110), requisitos (0.0100), ...	El tópico trata sobre Coahuila y su interacción con el gobierno federal, poniendo énfasis en concesiones y recursos federales. Incluye interrogantes sobre cantidades y detalles específicos, además de temas como estímulos económicos y contratos. Se mencionan proyectos de desarrollo urbano y municipal, con un enfoque en la prevención y manejo de desastres naturales.	Adquisiciones
2019	N.L.	8	Fiscalización y Estudios a Nivel Nacional	Nacional (0.1300), fiscal (0.1120), nivel (0.0300), transbordadores (0.0260), estudio (0.0200), número (0.0170), impugnación (0.0160), aludido (0.0160), institucional (0.0090), Mty (0.0090), ...	Centrado en la fiscalización y realización de estudios a nivel nacional, destaca la relevancia de los transbordadores y la necesidad de impugnación frente a decisiones aludidas. Se discute la estructura institucional y el papel de las naciones en el análisis fiscal, así como la importancia de los docentes y agremiados en el desarrollo y cooperación educativa.	Finanzas públicas
2018	Tamps.	0	Políticas Federales y Regulaciones en Salud y Educación	pensión (0.0430), federal (0.0320), estudio (0.0230), todo (0.0200), social (0.0140), presente (0.0110), documento (0.0110), López (0.0100), electoral (0.0100), legal (0.0100), ...	Este topic aborda las políticas federales relacionadas con pensiones, estudios sociales, y el marco legal electoral, destacando la importancia de la documentación y la regulación por parte de instituciones como COFEPRIS. Se enfoca en las iniciativas avaladas y los cursos implementados para la protección y el riesgo sanitario, subrayando la colaboración bilateral y las disposiciones financieras tomadas para el fondo federal.	Salud

5.4.4 Zona económica occidente

La Tabla 5.6 muestra la cantidad de tópicos correspondientes a la Zona Occidente, con un total de 391 tópicos durante el período analizado. Jalisco es el estado con la mayor frecuencia de tópicos, sumando 139 en total, seguido de Michoacán con 99, Nayarit con 82 y, finalmente, Colima con 71.

Tabla 5.6. Número de tópicos por estado en la zona económica occidente durante el periodo 2003–2020.

Estado	Número de Tópicos
Colima	71
Jalisco	139
Michoacán	99
Nayarit	82

La Figura 5.16, destaca la categoría de **Servidores Públicos**, con aumentos importantes en los años **2009, 2014, 2015 y especialmente 2017**, lo que indica un patrón de alta recurrencia en esta región. La categoría de **Medio Ambiente** mantiene una presencia constante a lo largo del tiempo, con mayor intensidad en **2011**, **2014** y **2017**. Por su parte, **Educación** y **Comercial** presentan picos en **2011**, **2014** y **2017**, lo que podría reflejar una atención especial hacia estos temas en dichos años. Las categorías de **Finanzas Públicas**, **Salud** y **Seguridad** muestran frecuencias moderadas pero estables, mientras que **Energía** y **Adquisiciones** presentan aumentos puntuales en años como **2012**, **2015** y **2019**.

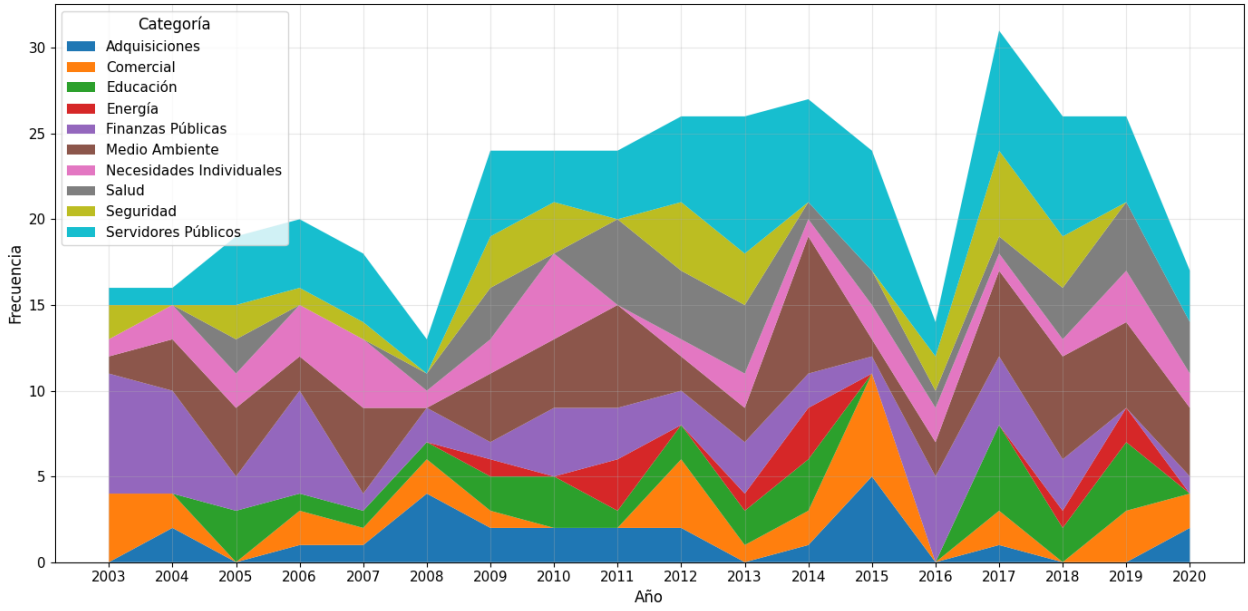

Figura 5.16. Tendencia de categorías durante el periodo 2003–2020 en la zona occidente.

Ejemplos de tópicos representativos zona occidente

La Tabla 5.7 muestra ejemplo de tópicos relevantes en la zona occidente de México, abarcando diversas áreas de interés en diferentes años y estados. En 2019, en Colima, el tópico 2 se enfoca en las becas y elecciones municipales, subrayando la educación y la participación ciudadana, así como la transparencia y la inclusión en procesos democráticos. En 2018, en Jalisco, el tópico 4 destaca la documentación y estudios ambientales realizados por SEMARNAT, enfatizando la importancia de registrar y comunicar adecuadamente los documentos relacionados con el medio ambiente. En el mismo año, en Michoacán, el tópico 2 analiza el financiamiento y costos de la infraestructura aeroportuaria, con un enfoque en el desarrollo económico y la gestión del tráfico aéreo. Finalmente, en 2013, en Nayarit, el tópico 1 se centra en los procesos federales y profesionales, incluyendo la gestión de transporte público y la interacción entre la administración pública y sectores específicos como el transporte y la defensa.Para más detalles de los años 2003 a 2019, consulte las Tablas en el Apéndice A.

Tabla 5.7. Tópicos representativos zona occidente

Año	Estado	Topic	Título	Probabilidades	Descripción	Categoría
2019	Col.	2	Becas y Elecciones Municipales Enfoque en la Educación y Participación Ciudadana	beca (0.0380), residente (0.0270), estudiante (0.0240), bully (0.0230), trabajadores (0.0210), elección (0.0190), dependencia (0.0190), coalición (0.0190), municipales (0.0190), suplente (0.0190), ...	Examina la asignación de becas para estudiantes y residentes, así como la participación en elecciones municipales y la coalición de trabajadores. Destaca la importancia de prevenir prácticas negativas como el bully y fomentar una educación integral. Se menciona el rol de las dependencias en registrar y actualizar información relevante para garantizar la transparencia y la participación efectiva del alumnado y los practicantes en procesos democráticos y educativos, promoviendo un enfoque proactivo hacia la inclusión y la diversidad en el ámbito municipal.	Educación
2018	Jal.	4	Documentación y Estudios Ambientales	Expediente (0.0210), general (0.0210), documento (0.0160), SEMARNAT (0.0140), federal (0.0140), estudio (0.0130), domicilio (0.0120), todo (0.0110), legal (0.0090), registrado (0.0080), ...	Aborda la importancia de la documentación legal y los estudios ambientales realizados por la SEMARNAT y otras entidades federales. Se destaca la necesidad de registrar y comunicar adecuadamente todos los documentos y estudios relacionados con el medio ambiente, recursos naturales y transporte, asegurando la adherencia a las disposiciones legales y reglamentarias.	Medio Ambiente
2018	Mich.	2	Infraestructura Aeroportuaria y su Financiamiento	número (0.0950), tasa (0.0350), entonces (0.0350), formato (0.0220), mayor (0.0150), periodo (0.0150), emanado (0.0150), gran ayuda (0.0140), licencias otorgadas (0.0140), referido (0.0130), ...	analiza el financiamiento y los costos asociados a la infraestructura de diversos aeropuertos, incluyendo el AICM y Toluca, enfocándose en los períodos y montos específicos para remodelar y adaptar estas instalaciones. Se destaca la importancia de desglosar estos gastos por conceptos y la implicación de estos proyectos en el desarrollo económico y la gestión de tráfico aéreo.	Energía
2013	Nay.	1	Procesos Federales y Profesionales	expediente (0.0410), federal (0.0210), otorgado (0.0120), público (0.0120), profesional (0.0120), certificado (0.0110), autobús (0.0100), obtenido (0.0090), quién (0.0080), mes (0.0070), ...	Trata sobre expedientes y certificaciones en contextos federales y profesionales, incluyendo la gestión de autobuses y transporte público. La referencia a militares retirados y evaluadores indica una interacción entre la administración pública y sectores específicos como el transporte y la defensa.	Servidores Públicos

5.4.5 Zona económica oriente

La Tabla 5.8 muestra la cantidad de tópicos identificados en la zona Económica Oriente, con un total de 425 tópicos. Puebla es el estado con el mayor número de tópicos, sumando 124 en total, seguido por Veracruz con 113, Hidalgo con 96 y, finalmente, Tlaxcala con 92.

Tabla 5.8. Número de tópicos por estado en la zona económica oriente durante el periodo 2003–2020.

Estado	**Número de Tópicos**
Hidalgo	96
Puebla	124
Tlaxcala	92
Veracruz	113

Con respecto a las categorías para la Zona Oriente, la Figura 5.17, se identifica un **pico general en 2015**, impulsado principalmente por las categorías de **Servidores Públicos**, **Seguridad**, **Salud**, **Necesidades Individuales** y **Medio Ambiente**, lo que sugiere un aumento transversal en temas de interés público. Por su parte, **Finanzas Públicas** presenta incrementos notables en los años **2012**, **2015** y **2019**, mientras que **Necesidades Individuales** mantiene una presencia constante. En contraste, categorías como **Adquisiciones**, **Comercial** y **Educación** muestran comportamientos más irregulares, sin una tendencia sostenida de crecimiento.

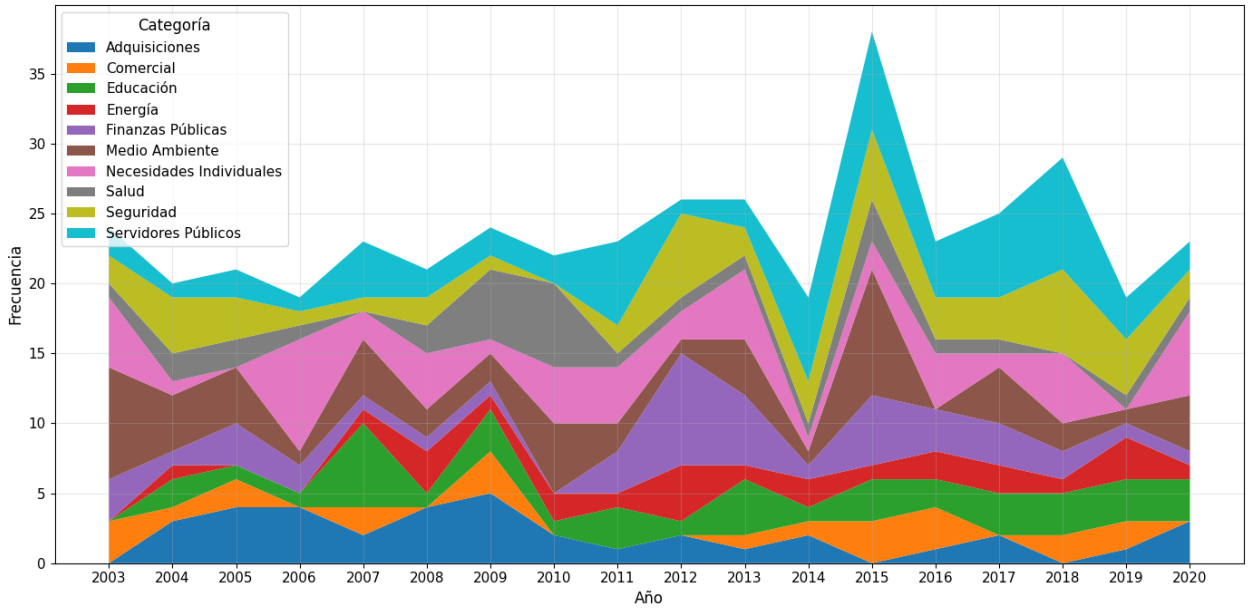

Figura 5.17. Tendencia de categorías durante el periodo 2003–2020 en la zona oriente.

Ejemplos de tópicos representativos zona oriente

La Tabla 5.9 muestra un ejemplo de tópicos representativos de los estados correspondientes a la zona oriente de México, abarcando diversos temas de interés en diferentes años. En 2006, en Hidalgo, el tópico 4 se centra en los recursos nacionales y el apoyo anual, incluyendo aspectos presupuestales y la administración rural, así como la prevención de enfermedades y la producción comunitaria. En 2013, en Puebla, el tópico 4 aborda la industria petrolera y sus efectos sobre la propiedad y las tierras, destacando la necesidad de explicaciones sobre procedimientos y criterios usados para determinar las afectaciones. En 2020, en Tlaxcala, el tópico 2 se enfoca en resoluciones penales y federativas en casos específicos como Puente Grande y Juanacatlán, subrayando la importancia de las decisiones judiciales y administrativas. Ese mismo año, en Veracruz, el tópico 6 examina el marco legal y las acciones en el contexto de la pandemia de COVID-19, resaltando la implementación de disposiciones legales para manejar la crisis sanitaria. Para más detalles de los años 2003 a 2019, consulte las Tablas en el Apéndice A.

Tabla 5.9. Tópicos representativos zona oriente

Año	Estado	Topic	Título	Probabilidades	Descripción	Categoría
2006	Hgo.	4	Recursos Nacionales y Apoyo Anual	recursos (0.0620), Nacional (0.0430), año (0.0420), apoyo (0.0370), sacar (0.0340), presupuestal (0.0290), consejo (0.0270), rubro (0.0270), servidores (0.0190), registrado (0.0170), ...	Este tópico se enfoca en los recursos nacionales y el apoyo anual. Incluye aspectos de presupuestos, consejos y rubros de servidores registrados. Trata sobre la construcción del CRIT en Ixmiquilpan y su funcionamiento. Examina la realización y preparación de festivales como el Cervantino. También aborda la administración rural, la prevención de enfermedades y la producción en comunidades respectivas.	Finanzas públicas
2013	Pue.	4	Industria Petrolera y Afectaciones a la Propiedad	petróleo (0.0660), propietarios poseedore (0.0330), tierra (0.0330), explique (0.0330), correspondiente (0.0280), afectó (0.0270), razones (0.0270), referente (0.0160), barrile (0.0150), periodo (0.0140), ...	Este tópico trata sobre la industria petrolera y cómo sus actividades afectan a las propiedades y tierras de los ciudadanos. Se solicita explicación sobre los procedimientos correspondientes, los criterios usados para determinar afectaciones, y las razones detrás de las decisiones tomadas. Los ciudadanos también buscan información sobre los volúmenes de producción y cómo estos impactan en el entorno y en la economía local.	Medio Ambiente
2020	Tlax.	2	Resoluciones Penales y Federativas en Puente Grande y Juanacatlán	Penal (0.0480), federales (0.0370), causa (0.0280), puente grande (0.0220), Juanacatlán (0.0220), resoluciones (0.0190), sobreseimiento (0.0150), antecede (0.0130), generales (0.0130), centro (0.0070), ...	Centrándose en el ámbito penal y federal, detalla las causas y resoluciones en casos específicos de Puente Grande y Juanacatlán, con una mención a la gestión de sobreseimientos y antecedentes generales. La referencia a centros, sedes y sistemas judiciales implica un enfoque en la infraestructura legal y carcelaria de México, subrayando la importancia de las decisiones judiciales y administrativas en el contexto de la justicia penal federal.	Seguridad
2020	Ver.	6	Análisis del Marco Legal y Acciones en el Contexto de COVID y Derechos	Presente (0.0900), artículo (0.0790), caso (0.0450), accione (0.0440), derecho (0.0430), covid (0.0310), análisis (0.0210), disposiciones (0.0190), administrativos (0.0180), realice (0.0170), ...	El Tópico explora el marco legal y las acciones pertinentes en el contexto de la pandemia de COVID y la protección de derechos. La presencia de términos como artículo, caso, acciones, y derecho resalta la importancia de las disposiciones legales y las medidas tomadas en respuesta a la pandemia. Se enfoca en el análisis y la implementación de acciones y disposiciones administrativas para manejar la crisis sanitaria, destacando el papel del derecho y las regulaciones en la gestión de la pandemia.	Necesidades Individuales

5.4.6 Zona económica centronorte

La Tabla 5.10 muestra la cantidad de tópicos identificados en la Zona Centro-Norte durante 2003-2020, con un total de 485 tópicos. Guanajuato es el estado con la mayor cantidad de tópicos identificados, sumando 111 en total. Le siguen San Luis Potosí con 108, Querétaro con 97, Zacatecas con 90 y, finalmente, Aguascalientes con 79.

Tabla 5.10. Número de tópicos por estado en la zona económica Centro-Norte durante el periodo 2003–2020.

Estado	Número de Tópicos
Aguascalientes	79
Guanajuato	111
Querétaro	97
San Luis Potosí	108
Zacatecas	90

La Figura 5.18, muestra un **pico general en 2015**, impulsado principalmente por las categorías de **Servidores Públicos**, **Seguridad**, **Salud** y **Necesidades Individuales**, lo cual sugiere un interés ampliado en áreas clave de gobernanza y derechos. A lo largo del periodo, **Servidores Públicos** mantiene una presencia constante con repuntes adicionales en **2008**, **2011**, **2014** y **2018**. La categoría de **Educación** también muestra aumentos importantes en **2011** y **2014**, mientras que **Medio Ambiente** y **Finanzas Públicas** reflejan patrones de crecimiento intermitente. En contraste, las categorías de **Comercial** y **Adquisiciones** presentan variaciones más irregulares a lo largo del tiempo.

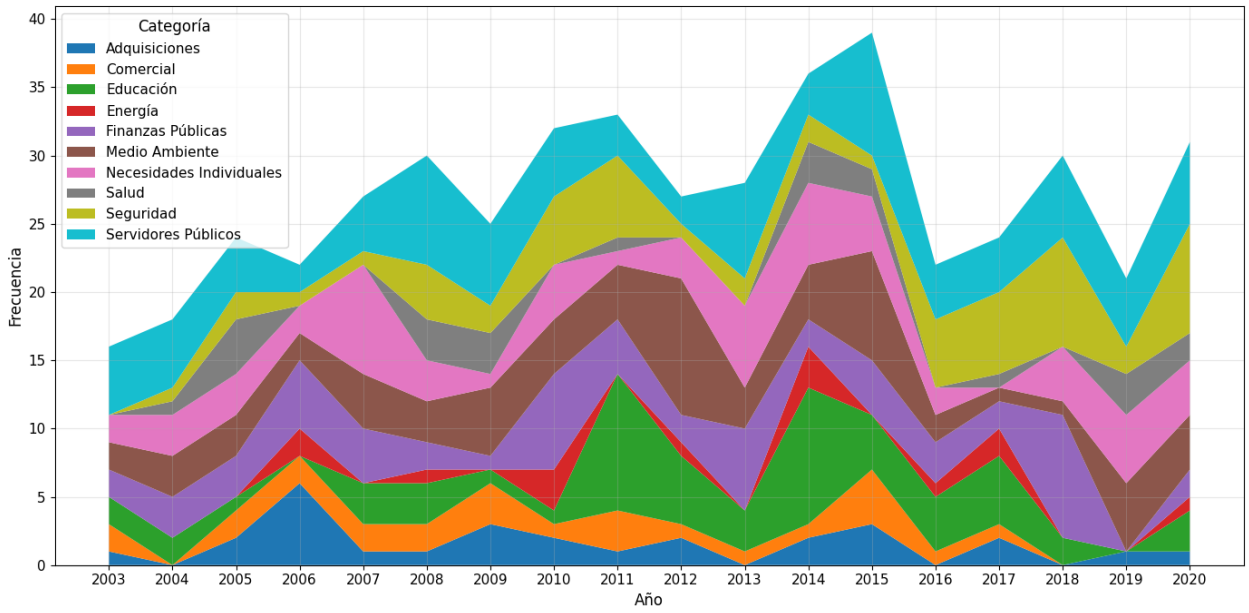

Figura 5.18. Tendencia de categorías durante el periodo 2003–2020 en la zona centro-norte.

Ejemplos de tópicos representativos zona centronorte

La Tabla 5.11 muestra un ejemplo de tópicos representativos de los estados correspondientes a la zona centronorte de México, abarcando distintos años y áreas temáticas. En 2006, en Aguascalientes, el tópico 3 se enfoca en concesiones y derechos a nivel nacional, abordando aspectos como la ampliación de empleados, verificaciones de aeropuertos, y auditorías fiscales. En 2009, en Guanajuato, el tópico 1 trata sobre análisis y patentes en el sector público, cubriendo investigaciones avanzadas y la relación entre tecnología y patentes, con mención de áreas específicas como Aguascalientes y Querétaro. En 2011, en Querétaro, el tópico 0 se centra en la gestión de recursos y adjudicaciones, discutiendo la responsabilidad en trámites y manejo de recursos del FOVISSSTE, además de aspectos educativos y profesionales. Finalmente, en 2015, en Zacatecas, el tópico 7 presenta los resultados de inspecciones y evaluaciones docentes, subrayando la importancia de la transparencia y el rendimiento educativo, así como la gestión de concursos y evaluaciones en el contexto educativo. Para más detalles de los años 2003 a 2019, consulte las Tablas en el apéndice.

Tabla 5.11. Tópicos representativos zona centronorte

Año	Estado	Topic	Título	Probabilidades	Descripción	Categoría
2006	Ags.	3	Concesiones y Derechos en el Ámbito Nacional	Ags. (0.0300), Nacional (0.0200), mensual (0.0200), instrucciones (0.0110), CNA (0.0110), hasta (0.0100), derechos (0.0100), nivel (0.0100), solicitar (0.0100), otorgado (0.0090), ...	Este tópico se enfoca en concesiones y derechos a nivel nacional, abordando instrucciones y situaciones en la CNA. Trata sobre ampliaciones, empleados y seguros, así como verificaciones y disponibilidad de aeropuertos. Incluye aspectos de mercado, fiscalización y auditorías, y la participación en diferentes categorías y funciones gubernamentales. También menciona incentivos agrarios y auditorías fiscales.	Adquisiciones
2009	Gto.	1	Análisis y Patentes en el Sector Público	análisis (0.0850), pública gubernamental (0.0840), patent (0.0830), negado (0.0820), número (0.0660), investigación (0.0120), indicadores (0.0060), investigaciones (0.0050), materiales (0.0050), sistema (0.0040), ...	Aborda el tema de análisis y patentes en el ámbito público y gubernamental. Se centra en investigaciones avanzadas y el uso de indicadores y materiales en el sistema de transporte. También cubre áreas geográficas específicas como Aguascalientes y Querétaro, y la relación entre tecnología y tramitación de patentes.	Comercial
2011	Qro.	0	Gestión de Recursos y Adjudicaciones en QRO	número (0.0220), recursos (0.0180), responsable (0.0120), aclaraciones (0.0100), archivos (0.0100), adjudicación (0.0100), QRO (0.0100), tiempo (0.0090), FOVISSSTE (0.0090), transferir (0.0090), ...	Este tópico se centra en la gestión de recursos y procesos de adjudicaciones en Querétaro (QRO). Incluye la responsabilidad y aclaraciones sobre archivos y trámites, y el manejo de recursos del FOVISSSTE. Se discute sobre la carga y descarga en terminales y domicilios, y los aspectos profesionales y fiscales involucrados. Se mencionan también temas educativos, inventarios y servicios relacionados con la salud y la industria, con un enfoque en la representación sindical y asambleas.	Adquisiciones
2015	Zac.	7	Resultados de Inspecciones y Evaluaciones Docentes	obtenido (0.1030), solo (0.1010), periodo (0.0830), inspectores (0.0720), docente (0.0670), resultado (0.0610), inspector (0.0200), organigrama (0.0150), nepotismo (0.0150), publicación (0.0150), ...	Aquí se presentan los resultados de inspecciones y evaluaciones docentes. Se destaca la importancia de la transparencia y el rendimiento educativo, incluyendo detalles sobre inspecciones, nepotismo y concursos. Además, se discuten los aspectos relevantes para el desarrollo educativo, tales como el ciclo escolar y los procesos de evaluación docente.	Educación

5.4.7 Zona económica centrosur

La Tabla 5.12 muestra la cantidad de tópicos identificados en la Zona Centro-Sur durante el período 2003-2020, con un total de 1158 tópicos. La Ciudad de México es la entidad con el mayor número de tópicos, sumando 788 en total, seguida por el Estado de México con 273 tópicos y, finalmente, Morelos con 97 tópicos.

Tabla 5.12. Número de tópicos por estado en la zona económica Centro-Sur durante el periodo 2003–2020.

Estado	**Número de Tópicos**
Ciudad de México	788
Morelos	97
México	273

La Figura 5.19 muestra un **pico general en 2017**, impulsado principalmente por las categorías de **Servidores Públicos**, **Seguridad**, **Salud** y **Necesidades Individuales**, lo que refleja una fuerte concentración de solicitudes en áreas institucionales y sociales clave. La categoría de **Servidores Públicos** destaca por su alta frecuencia sostenida en el tiempo, con repuntes notables en **2010**, **2013** y **2017**. Por su parte, **Seguridad** muestra una tendencia creciente a partir de 2010, alcanzando su punto más alto en 2017, y **Salud** presenta un aumento constante desde 2013. En contraste, categorías como **Medio Ambiente**, **Comercial** y **Energía** mantienen niveles bajos o intermitentes a lo largo del periodo.

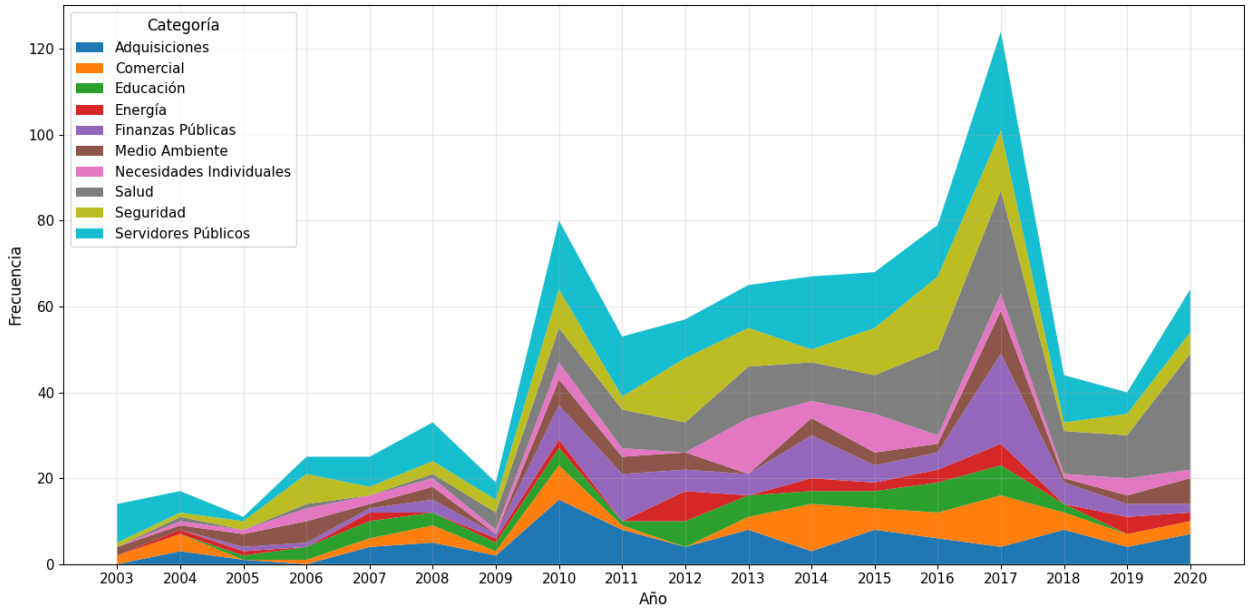

Figura 5.19. Tendencia de categorías durante el periodo 2003–2020 en la zona centro-sur.

Ejemplos de tópicos representativos zona centrosur

La Tabla 5.13 muestra un ejemplo de tópicos representativos de los estados correspondientes a la zona centrosur de México. En 2004, en la Ciudad de México, el tópico 10 se enfoca en el cumplimiento de contratos y la gestión de órdenes diversas, resaltando la supervisión de incumplimientos y la importancia de mantener registros actualizados en el IMSS y otros laboratorios. En 2014, en el Estado de México, el tópico 2 aborda la transparencia en las funciones federales, destacando el uso de portales para la facturación y la delegación de funciones, subrayando un compromiso con la rendición de cuentas. En 2015, en Morelos, el tópico 1 trata sobre la protección de derechos y la transparencia, destacando la importancia de proporcionar copias y documentos notariales para garantizar la claridad en las acciones administrativas. Para más detalles de los años 2003 a 2019, consulte las Tablas en el Apéndice A.

Tabla 5.13. Tópicos representativos zona centrosur

Año	Estado	Topic	Título	Probabilidades	Descripción	Categoría
2004	CDMX	10	Cumplimiento de Contratos y Ordenes Diversas	todo (0.0550), contrato (0.0380), diversas órdenes (0.0290), incurrió (0.0290), incumplimientos (0.0290), IMSS delegación (0.0170), lab (0.0170), mediante (0.0170), fiscal (0.0120), disponible (0.0110), ...	El tópico trata sobre el cumplimiento de contratos y la gestión de órdenes diversas. Se enfoca en la supervisión de incumplimientos y la importancia de mantener registros actualizados en el IMSS y otros laboratorios. Se examina el papel de los proveedores y las penalizaciones derivadas del incumplimiento de contratos. Además, se consideran los manuales y direcciones generales involucradas en estos procesos, resaltando la importancia de una revisión y gestión eficaz.	Adquisiciones
2014	Méx.	2	Transparencia y Funciones Federales	federal (0.0440), funciones (0.0260), obligaciones (0.0230), portal (0.0210), factura (0.0150), sede (0.0140), resultado (0.0130), cargo (0.0110), detallado (0.0110), delegación (0.0090), ...	Este tópico examina la transparencia en las funciones federales, destacando la importancia del portal federal para la facturación y los resultados de las funciones de cada cargo. La delegación de funciones y la inspección anual son temas clave, indicando un esfuerzo por evitar conflictos de interés y mejorar la integridad en la administración pública. Se menciona el detalle de las obligaciones y los mecanismos para su cumplimiento, reflejando un compromiso con la rendición de cuentas y la gestión eficaz.	Servidores Públicos
2015	Mor.	1	Protección de Derechos y Transparencia	tiempo (0.0250), marca (0.0150), copia (0.0090), derechos impuesto (0.0090), notarial (0.0090), proporcione (0.0080), tema (0.0060), general (0.0060), vivienda (0.0060), recibos oficiales (0.0060), ...	Trata sobre la importancia de proporcionar copias y documentos notariales para garantizar la transparencia y protección de derechos. Este tópico resalta la necesidad de cumplir con requisitos específicos y proporcionar evidencia clara de las acciones y decisiones tomadas, subrayando la importancia de la transparencia en la administración pública.	Necesidades Individuales

5.4.8 Zona económica suroeste

La Tabla 5.14 muestra la cantidad de tópicos identificados en la Zona Suroeste durante el período 2003-2020, con un total de 285 tópicos. Chiapas es el estado con la mayor frecuencia de tópicos identificados, con un total de 101, seguido por Guerrero con 93 y Oaxaca con 91.

Tabla 5.14. Número de tópicos por estado en la zona económica Suroeste durante el periodo 2003–2020.

Estado	**Número de Tópicos**
Chiapas	101
Guerrero	93
Oaxaca	91

La Figura 5.20 muestra un **pico general en 2015**, impulsado principalmente por las categorías de **Servidores Públicos**, **Seguridad**, **Salud** y **Necesidades Individuales**, lo que indica una alta concentración de solicitudes relacionadas con aspectos administrativos y sociales. A lo largo del periodo, la categoría de **Servidores Públicos** mantiene una frecuencia alta y constante, con repuntes en **2010**, **2014** y nuevamente en **2020**. **Seguridad** muestra una tendencia ascendente a partir de 2010, con niveles elevados en 2014 y entre 2018 y 2020. En contraste, **Medio Ambiente**, **Comercial** y **Adquisiciones** presentan una participación más irregular y limitada a lo largo del periodo.

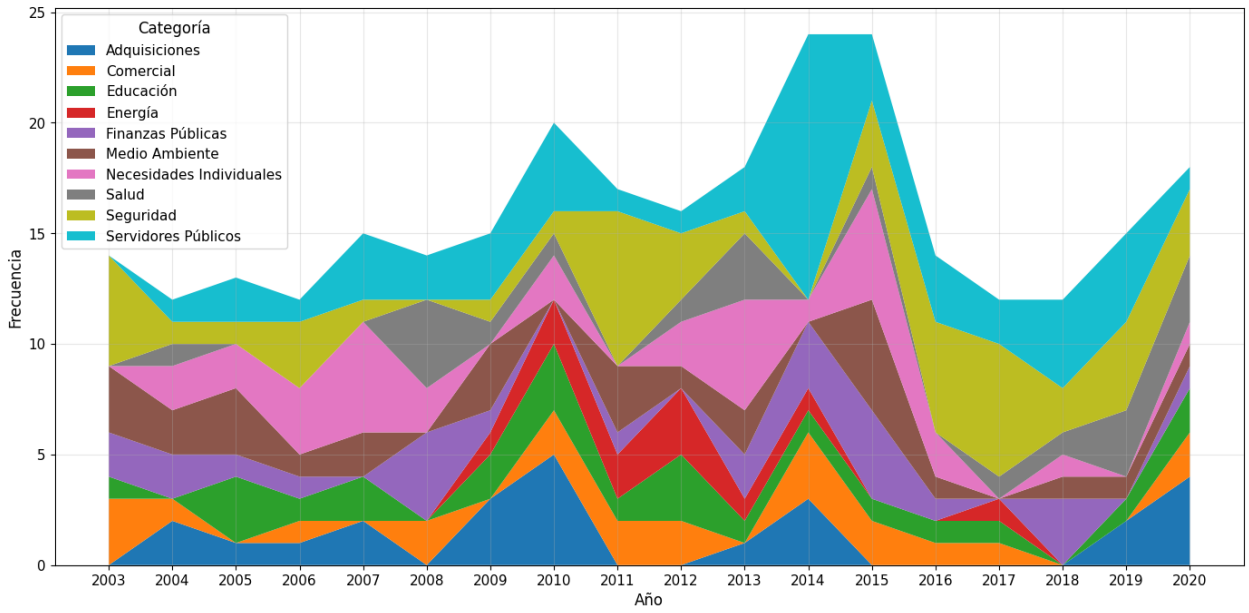

Figura 5.20. Tendencia de categorías durante el periodo 2003–2020 en la zona suroeste.

Ejemplos de tópicos representativos zona suroeste

La Tabla 5.15 muestra un ejemplo de tópicos de los estados correspondientes a la zona suroeste de México, abarcando diferentes años y temas clave. En 2014, en Chiapas, el tópico 3 se enfoca en los procesos de licitación y responsabilidad, subrayando la importancia de cumplir con criterios técnicos y legales en la selección de licitantes para servicios gubernamentales. En 2004, en Guerrero, el tópico 0 aborda la administración agraria a nivel nacional, incluyendo la gestión de superficies de terrenos y precios de materiales de construcción, así como aspectos fiscales relacionados con el sector agrario. En 2009, en Oaxaca, el tópico 1 se centra en los contratos laborales de profesores en escuelas particulares y campos experimentales, destacando los contratos individuales y de tiempo completo, así como las prestaciones y el impacto de prácticas agrícolas en el desarrollo profesional y académico. Para más detalles de los años 2003 a 2019, consulte las Tablas en el Apéndice A.

Tabla 5.15. Tópicos representativos zona suroeste

Año	Estado	Topic	Título	Probabilidades	Descripción	Categoría
2014	Chis.	3	Procesos de Licitación y Responsabilidad	caso (0.0370), responsable (0.0360), los puntos (0.0320), licitante (0.0320), firmado por (0.0320), dictamen técnico (0.0320), desecharon expresando (0.0320), razones legales (0.0320), convocatoria (0.0320), técnicas económicas (0.0320), ...	Trata sobre los procesos de licitación y la responsabilidad en la selección de licitantes, enfatizando la importancia de cumplir con los criterios técnicos y legales. Palabras clave como caso, responsable, y licitante sugieren un análisis detallado de casos específicos donde se evaluaron propuestas técnicas y económicas para servicios gubernamentales.	Adquisiciones
2004	Gro.	0	Administración Nacional Agraria y Construcción	Nacional (0.0360), agrario (0.0170), superficie (0.0170), CURP (0.0160), varilla (0.0140), loza (0.0140), precio (0.0110), impuesto (0.0100), reintegro (0.0100), individual (0.0100), ...	Este tópico se enfoca en la administración agraria a nivel nacional, incluyendo temas como la superficie de terrenos, el uso de la CURP en trámites agrarios, y precios de materiales de construcción como varillas y lozas. Aborda aspectos fiscales como impuestos y reintegros en el sector agrario, y la importancia de la individualidad y precisión en las peticiones y resolución de inconformidades.	Medio Ambiente
2009	Oax.	1	Contratos Laborales de Profesores en Escuelas Particulares y Campos Experimentales	profesor (0.0840), contrato (0.0390), individual (0.0370), íntegro (0.0360), tiempo completo (0.0360), contratos laborales (0.0360), escuelas particulares (0.0340), positiva (0.0340), campos experimentales (0.0220), clase (0.0190), ...	El tópico trata sobre contratos laborales de profesores, enfocándose en contratos individuales y de tiempo completo en escuelas particulares. Se examina la naturaleza positiva y el ámbito de campos experimentales relacionados con la educación. Se discute sobre prestaciones, fechas anuales y diferentes tipos de contratos laborales. También se toca el tema de semillas genéticamente modificadas, su cosecha y el impacto de estas prácticas en el desarrollo profesional y académico.	Educación

5.4.9 Zona económica sureste

La Tabla 5.16 muestra la cantidad de tópicos identificados en la Zona Sureste durante el período 2003-2020, con un total de 386 tópicos. Yucatán es el estado con el mayor número de tópicos identificados, sumando 110 en total, seguido por Tabasco con 100, Quintana Roo con 97 y Campeche con 79.

Tabla 5.16. Número de tópicos por estado en la zona económica Sureste durante el periodo 2003–2020.

Estado	**Número de Tópicos**
Campeche	79
Quintana Roo	97
Tabasco	100
Yucatán	110

La Figura 5.21 muestra un **pico general en 2018**, impulsado principalmente por las categorías de **Servidores Públicos**, **Seguridad**, **Salud** y **Necesidades Individuales**, lo que sugiere una atención creciente a temas institucionales y sociales. A lo largo del periodo, la categoría de **Medio Ambiente** mantiene una presencia elevada y constante, con máximos visibles en **2010**, **2015** y **2018**. Por su parte, **Servidores Públicos** muestra una evolución ascendente entre 2003 y 2016, seguida de una caída abrupta en 2019. **Educación** y **Finanzas Públicas** presentan aumentos significativos en **2009**, **2014** y **2018**, mientras que otras categorías como **Comercial** muestran una participación más irregular y de menor intensidad.

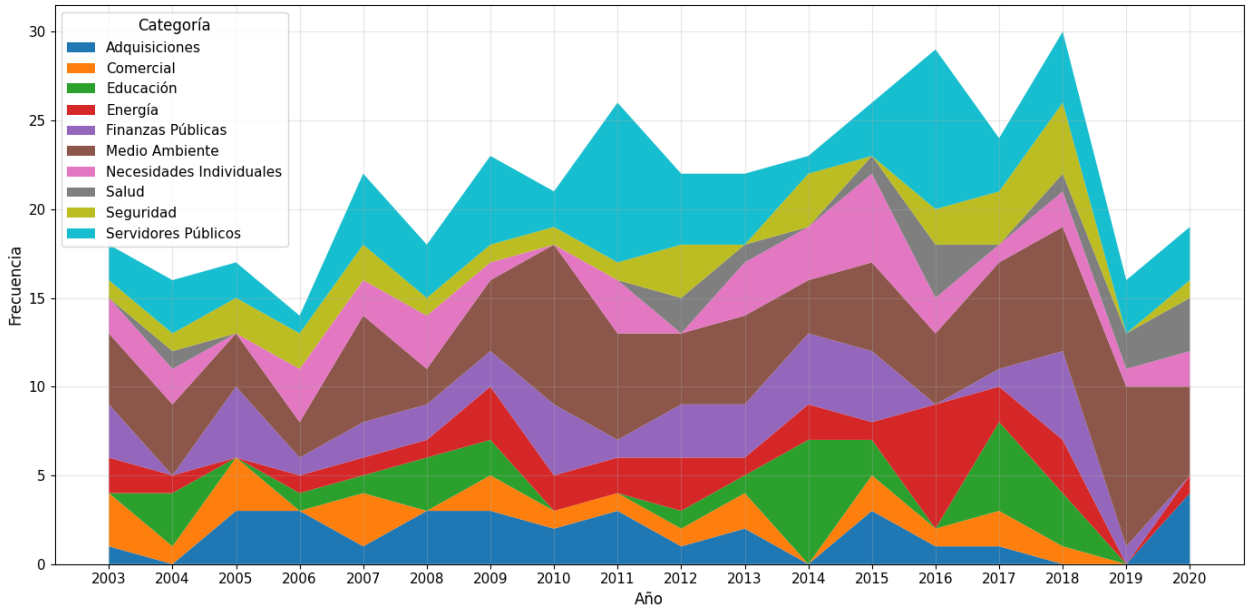

Figura 5.21. Tendencia de categorías durante el periodo 2003–2020 en la zona sureste.

Ejemplos de tópicos representativos zona sureste

La Tabla 5.17 muestra un ejemplo de tópicos de los estados correspondientes a la zona sureste de México, abarcando distintos años y áreas de interés. En 2011, en Campeche, el tópico 3 se enfoca en el desarrollo y promoción de proyectos empresariales, incluyendo la infraestructura y la participación empresarial en proyectos de fibra óptica y turísticos. Ese mismo año, en Quintana Roo, el tópico 5 trata sobre la producción y fiscalización de productos energéticos y marinos, abordando aspectos como el diésel, el asfalto y la tala ilegal.

En 2014, en Tabasco, el tópico 2 se centra en la administración de recursos a nivel federal y municipal, destacando la transparencia y la gestión de recursos en hospitales regionales. Finalmente, en 2020, en Yucatán, el tópico 2 examina las responsabilidades y normativas para trabajadores en organismos públicos como SENEAM, enfatizando la importancia de los informes, la defensa de derechos y el cumplimiento de normativas vigentes. Para más detalles de los años 2003 a 2019, consulte las Tablas en el Apéndice A.

Tabla 5.17. Tópicos representativos zona sureste

Año	Estado	Topic	Título	Probabilidades	Descripción	Categoría
2011	Camp.	3	Desarrollo y Promoción de Proyectos Empresariales	proyecto (0.0060), poste (0.0060), colocar (0.0060), empresa (0.0050), fibra (0.0050), promoción (0.0040), boleta (0.0040), trámite (0.0040), través (0.0040), carácter (0.0040), ...	Se centra en el desarrollo y promoción de proyectos empresariales, incluyendo la colocación de infraestructuras como postes y la participación de empresas en la fibra óptica. Aborda los trámites legales y riesgos asociados, así como la promoción en Playa Esmeralda y otros lugares. Se destaca la importancia de la representación legal y las oportunidades de desarrollo turístico y construcción.	Comercial
2011	Q.Roo	5	Producción y Fiscalización de Productos Energéticos y Marinos	volumen (0.0080), especial (0.0070), marino (0.0070), diésel (0.0070), anual (0.0060), producción (0.0050), importaciones (0.0040), asfalto (0.0040), precio (0.0040), fiscal (0.0040), ...	Este tópico analiza la producción y fiscalización anual de productos energéticos y marinos, incluyendo diesel, asfalto y especies marinas. Se enfoca en precios, fiscalidad, importaciones y exportaciones, así como en el reciclado de productos. Aborda temas municipales, domicilios, especies maderables y la tala ilegal. Incluye aspectos relacionados con la investigación pericial y la función de los grupos en la producción energética.	Energía
2014	Tab.	2	Administración de Recursos y Gobierno Municipal	recursos (0.0310), todo (0.0200), federal (0.0200), informe (0.0170), tiempo (0.0130), ocurrido (0.0130), gobierno (0.0110), municipal (0.0110), civil (0.0090), personales (0.0080), ...	Este tópico aborda la administración de recursos a nivel federal y municipal, resaltando la importancia de la transparencia y el acceso a la información gubernamental. Se menciona la gestión de personal, incidentes específicos y el resguardo de recursos en hospitales regionales, subrayando la interacción entre el gobierno y la administración local en el manejo de recursos y servicios civiles.	Finanzas públicas
2020	Yuc.	2	Responsabilidades y Normativas para Trabajadores en Organismos Públicos	Trabajadores (0.0480), seneam (0.0190), organismo (0.0120), mexicano (0.0110), hacienda crédito (0.0100), secretaría (0.0100), informe (0.0090), entregado (0.0080), establecido (0.0070), defensa (0.0070), ...	Este tópico examina las obligaciones y normativas específicas para trabajadores dentro de organismos como SENEAM, enfocándose en la Secretaría de Hacienda y Crédito Público. Detalla el proceso de informes entregados, la defensa de derechos de los contribuyentes, y las bases establecidas para la elaboración de manuales y guías de procedimiento. Subraya la importancia de responder prontamente a las inquietudes de los trabajadores, respetando los niveles salariales y sociales establecidos, y destaca la necesidad de especificar funciones y gastos conforme a las normativas vigentes.	Servidores Públicos

Una etapa clave de esta investigación fue la clasificación de los tópicos, tomando como referencia las categorías propuestas por Berliner et al. (2022) para facilitar la comparación con estudios previos. La clasificación se realizó manualmente, a partir del análisis cualitativo de las

palabras clave y sus probabilidades en cada tópico generado por el modelo LDA. Si bien las categorías generales fueron útiles, se identificó una riqueza temática que sugiere la necesidad de una mayor granularidad. Por ejemplo, dentro de la categoría *Medio Ambiente* emergieron subtemas como sustentabilidad, legislación ambiental, administración del agua y biodiversidad, que podrían abordarse de forma más específica.

La Figura 5.22 compara nuestros resultados (izquierda) con los obtenidos por Berliner et al. (2022) (derecha). En nuestro análisis, la categoría de **Servidores Públicos** presenta una mayor variabilidad temporal, reflejando cambios más marcados en la atención institucional a lo largo del tiempo. En contraste, el estudio de Berliner muestra una dominancia más estable de la categoría de **Finanzas Públicas**. Ambos estudios coinciden en la detección de un pico en la categoría **Comercial** durante el año 2013.

Aunque ambos enfoques utilizan modelado de tópicos con LDA, difieren en aspectos clave como la selección del número de tópicos y los criterios de clasificación. Nuestro modelo optimizó automáticamente el número de tópicos mediante un algoritmo evolutivo basado en coherencia temática, complementado con una clasificación manual experta. Esta estrategia permitió captar una mayor variabilidad interanual y una representación más matizada de los temas.

En conjunto, la comparación evidencia cómo la combinación de técnicas automatizadas con análisis interpretativo mejora la capacidad de capturar cambios temáticos relevantes en las solicitudes de información pública.

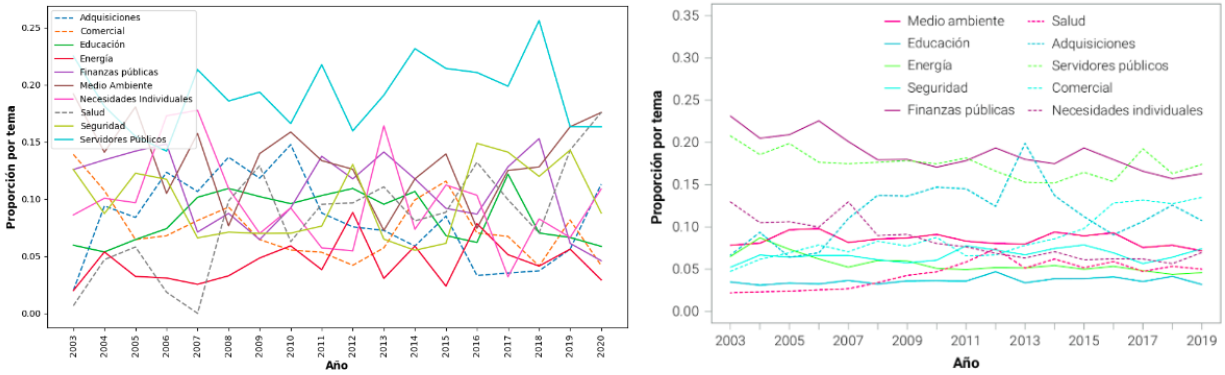

Figura 5.22. Comparación de categorías: clasificación manual vs. Berliner et al. (2022). **Fuente:** La parte izquierda fue elaborada por el autor, mientras que la derecha corresponde a Berliner et al. (2022).

5.5 Clasificación automática de tópicos

En este apartado se presentan los resultados obtenidos mediante el uso de clasificadores automáticos de tópicos basados en los modelos GPT, LLaMA y BETO, aplicados a la clasificación de tópicos relacionados con el acceso a la información pública gubernamental, según las categorías propuestas por (Berliner et al., 2022). Cada uno de estos modelos ha sido evaluado en su capacidad para identificar y categorizar correctamente la información, considerando tanto su precisión como su eficacia en el manejo de la complejidad y diversidad de los datos. Los resultados se analizan en detalle para destacar las fortalezas y limitaciones de cada enfoque, proporcionando una visión integral de cómo estos modelos pueden facilitar el análisis automatizado de datos textuales en el ámbito del acceso a la información pública.

La Figura 5.23 presenta los resultados de un clasificador basado en GPT que utiliza la técnica de Zero-Shot Prompting. Este enfoque permite al modelo clasificar automáticamente los tópicos en una de las diez categorías propuestas, basándose únicamente en el título y la descripción de cada tópico, sin requerir un entrenamiento previo específico para las categorías. La matriz de confusión proporciona un análisis detallado del rendimiento del clasificador para cada categoría.

En cuanto a la categoría de Servidores Públicos, destaca por su alto número de aciertos (764), lo que indica que el modelo identifica con gran precisión y recall los tópicos pertenecientes a esta clase, aunque existen algunas confusiones menores con las categorías de Finanzas Públicas y Educación, sugiriendo cierta ambigüedad en casos puntuales. Las categorías de Medio Ambiente y Seguridad también muestran un rendimiento sólido, con 500 y 367 aciertos respectivamente, aunque se observan errores esporádicos que podrían estar relacionados con similitudes contextuales entre algunas clases. Por su parte, las categorías de Educación y Finanzas Públicas obtienen 344 y 426 aciertos respectivamente, aunque presentan algunas confusiones con otras clases, lo que sugiere que los títulos y descripciones en estas categorías podrían compartir características con otras. Las categorías de Energía y Salud presentan una mayor cantidad de errores de clasificación, probablemente debido a la falta de términos distintivos claros en los títulos y descripciones utilizados para estas clases, lo que indica que el modelo podría beneficiarse de ajustes adicionales en estas áreas.

Las métricas de recall y F1-score muestran que las categorías de Servidores Públicos y Medio Ambiente presentan una alta capacidad del modelo para recuperar tópicos relevantes dentro de estas clases. Además, el F1-score indica un equilibrio adecuado entre precisión y recall en las categorías de Servidores Públicos, Medio Ambiente y Finanzas Públicas, lo que refleja un buen desempeño general del modelo en la clasificación temática.

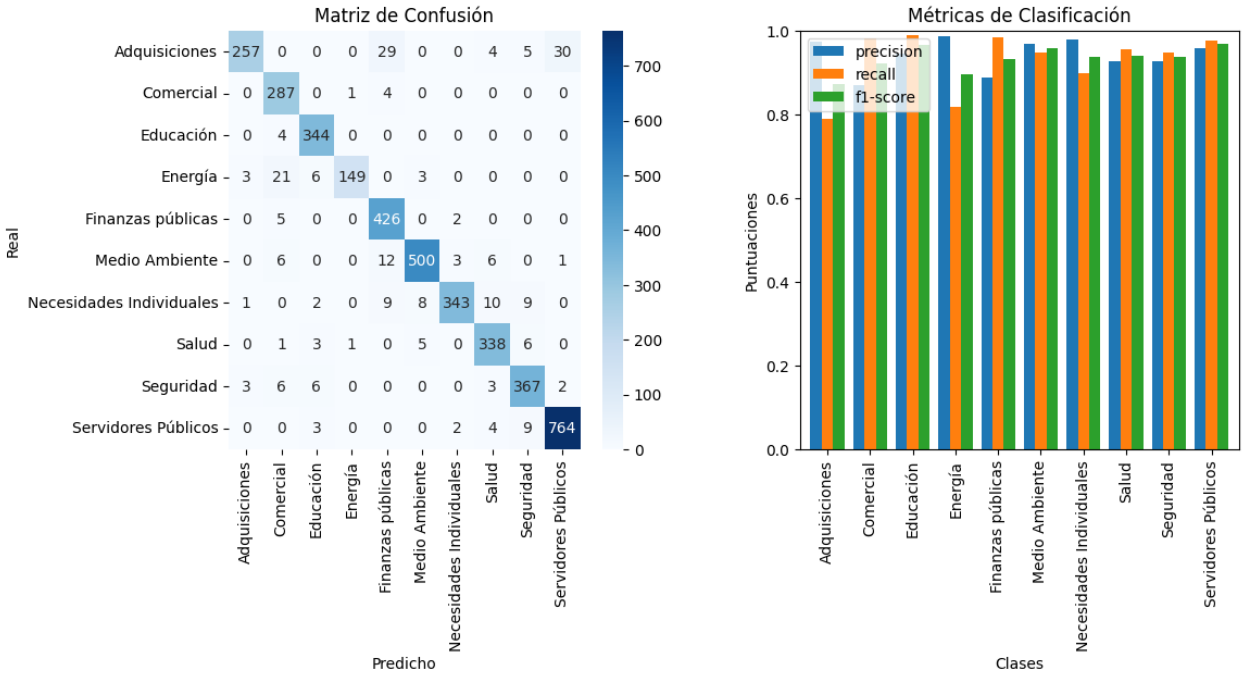

Figura 5.23. Matriz de confusión para el modelo GPT.

La Figura 5.24 muestra los resultados de un clasificador basado en LLaMA que utiliza la técnica de Zero-Shot Prompting.La matriz de confusión, que proporciona un análisis del rendimiento del clasificador para cada categoría.

La categoría de Adquisiciones destaca por su alto número de aciertos (310), lo que indica que el modelo es bastante efectivo en la clasificación de tópicos relacionados con esta área. Sin embargo, se observan ligeras confusiones con las categorías de Seguridad y Salud. Las categorías de Comercial y Educación también presentan un buen desempeño, con 261 y 273 aciertos respectivamente. No obstante, Educación muestra algunas confusiones con Medio Ambiente y Servidores Públicos, sugiriendo que ciertos tópicos podrían compartir características comunes que dificultan la clasificación precisa. Las categorías de Finanzas Públicas y Medio Ambiente tienen un número respetable de aciertos (361 y 381, respectivamente), aunque Finanzas Públicas muestra una mayor confusión con otras categorías como Servidores Públicos y Educación, lo que podría indicar que los tópicos en estas áreas comparten terminología o contextos similares.

Por otro lado, las categorías de Salud y Seguridad muestran resultados más variados. Salud registra 261 aciertos, pero también presenta una cantidad significativa de errores de clasificación, mientras que Seguridad alcanza 317 aciertos, con confusiones notables en las predicciones, especialmente con las categorías de Adquisiciones y Finanzas Públicas. En cuanto a Servidores Públicos, esta categoría presenta un rendimiento mixto, con 475 aciertos, pero también una notable cantidad de errores, al confundir tópicos con Finanzas Públicas y Educación.

En términos de precisión, las categorías de Adquisiciones y Medio Ambiente destacan por su alto desempeño, lo que indica que cuando el modelo clasifica un tópico en estas categorías, es muy probable que la clasificación sea correcta. En cuanto al recall, las categorías de Finanzas Públicas y Servidores Públicos muestran valores elevados, sugiriendo que el modelo es efectivo en capturar la mayoría de los tópicos relevantes en estas clases, aunque con una precisión que podría mejorar. El F1-score, que combina precisión y recall, muestra que las categorías de Medio Ambiente y Adquisiciones logran un buen equilibrio, mientras que Finanzas Públicas y Salud presentan un desempeño más irregular, lo que refleja la necesidad de ajustes adicionales en la clasificación de estos tópicos.

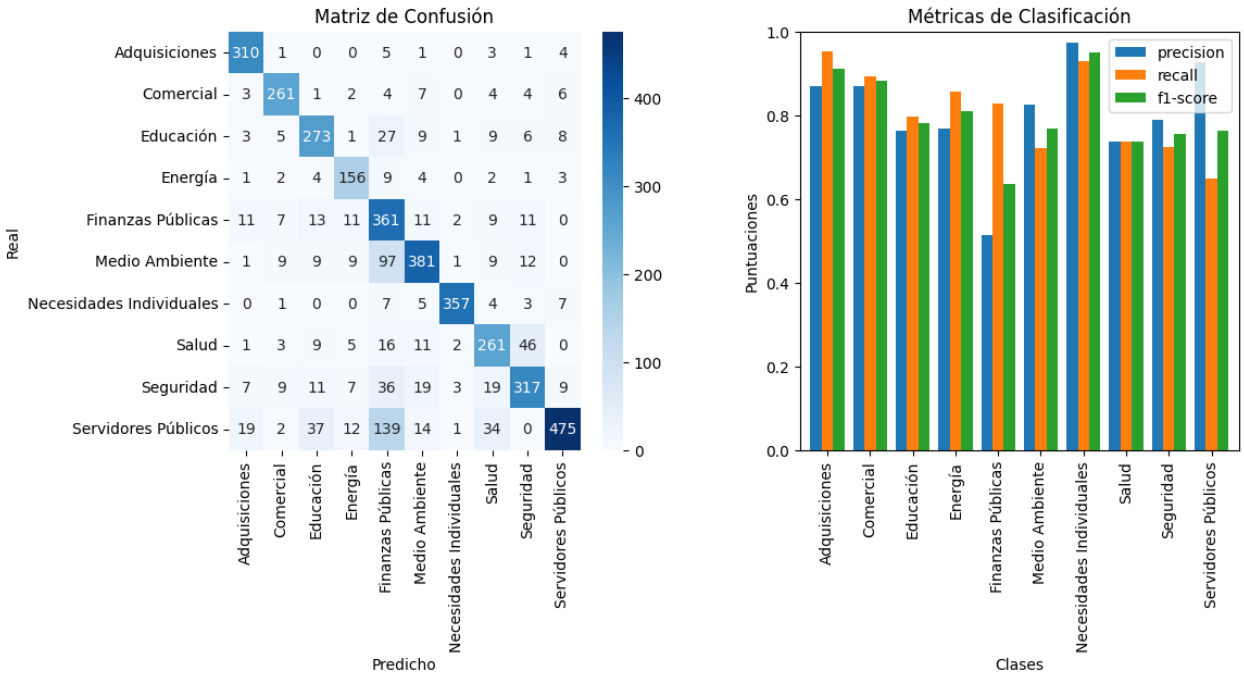

Figura 5.24. Matriz de confusión para el modelo LLaMA.

La Figura 5.25 muestra los resultados de un clasificador basado en BETO (dccuchile/bert-base-spanish-wwm-cased), un modelo de red neuronal tipo transformer entrenado específicamente para el idioma español. En la Figura de la izquierda se proporciona un análisis detallado del

rendimiento del clasificador para cada categoría. La categoría de Adquisiciones presenta un número considerable de aciertos (205), aunque también se observan confusiones significativas con otras categorías como Finanzas Públicas y Seguridad, lo que podría indicar que los tópicos relacionados con adquisiciones comparten términos o contextos con estas categorías. Las categorías de Comercial y Educación, con 185 y 288 aciertos respectivamente, muestran un rendimiento razonablemente bueno, aunque Comercial presenta confusiones notables con Finanzas Públicas y Energía, mientras que Educación se confunde ocasionalmente con Medio Ambiente y Necesidades Individuales.

Por su parte, las categorías de Finanzas Públicas y Medio Ambiente destacan con 304 y 429 aciertos respectivamente, pero también presentan cierta confusión con Servidores Públicos y otras categorías, lo que podría reflejar la complejidad de los tópicos en estas áreas, donde es difícil para el modelo separar claramente las categorías. La categoría de Salud, con 290 aciertos, muestra errores dispersos en varias otras categorías, sugiriendo que los tópicos relacionados con la salud pueden estar distribuidos en diversas áreas temáticas. Seguridad, con 297 aciertos, presenta confusiones especialmente con Adquisiciones y Finanzas Públicas, lo que indica una posible superposición en los términos utilizados en estos tópicos. Finalmente, la categoría de Servidores Públicos muestra un rendimiento mixto, con 423 aciertos, pero también una cantidad significativa de errores, particularmente al confundirse con Finanzas Públicas, Educación y Medio Ambiente, lo que sugiere que la distinción entre estas categorías no siempre es clara en los datos analizados.

Las categorías de Medio Ambiente y Educación destacan por su alta precisión, lo que indica que cuando el modelo clasifica un tópico en estas categorías, es muy probable que la clasificación sea correcta. Medio Ambiente y Salud muestran valores elevados de recall, lo que sugiere que el modelo es efectivo en capturar la mayoría de los tópicos relevantes en estas clases, aunque con una precisión que podría mejorar. El F1-score, que combina precisión y recall, muestra que las categorías de Medio Ambiente, Educación y Seguridad logran un buen equilibrio en su clasificación. Sin embargo, categorías como Adquisiciones y Necesidades Individuales muestran un desempeño más irregular, lo que refleja la necesidad de ajustes adicionales en la clasificación de estos tópicos.

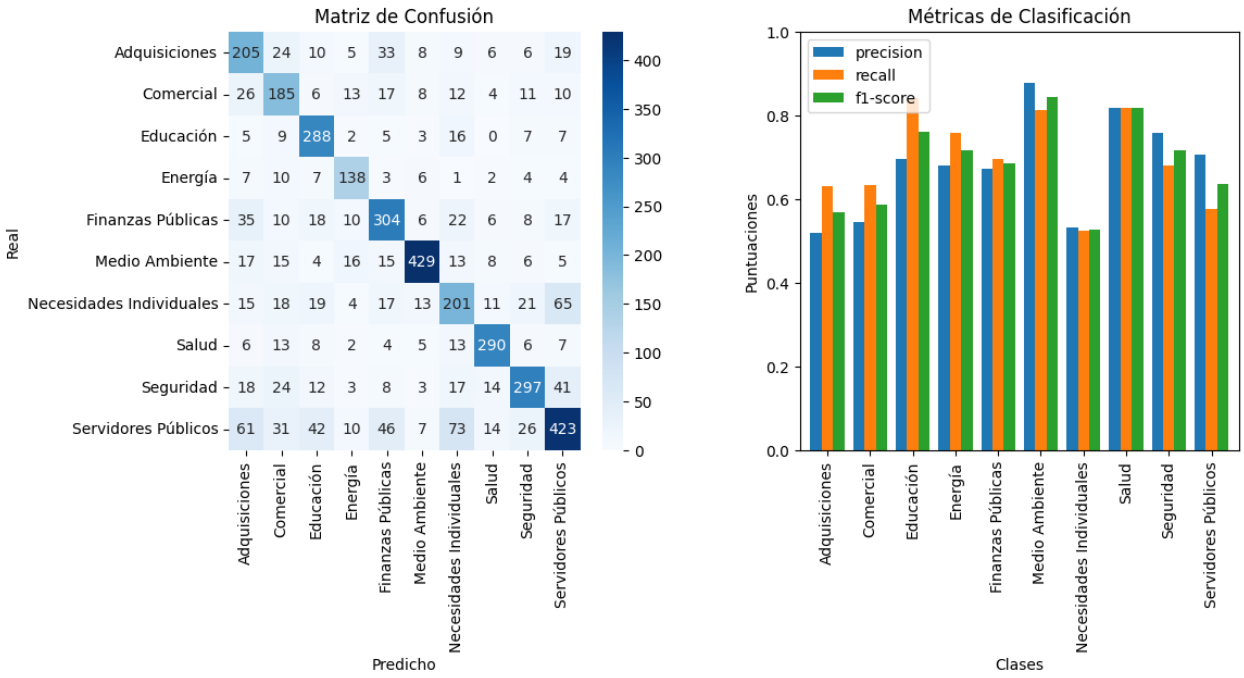

Figura 5.25. Matriz de confusión para el modelo BETO.

5.6 Resultados del análisis de solicitudes en el sector ambiental

Con respecto al análisis de las solicitudes de información ambiental, se llevó a cabo un estudio del sector medioambiental en México. La metodología empleada permitió identificar 115 tópicos clave, revelando tendencias en el interés público a lo largo del tiempo. La Figura 5.26 presenta la variabilidad temporal en la frecuencia de discusión de tópicos medioambientales, clasificados en cinco categorías principales, entre 2003 y 2020. Los resultados muestran fluctuaciones notables que podrían estar relacionadas con cambios sociopolíticos, la promulgación de nuevas legislaciones ambientales y la ocurrencia de eventos críticos, como desastres naturales o la implementación de políticas públicas relevantes.

Se observa un **pico general entre 2006 y 2009**, con una mayor concentración de temas relacionados con **Gestión de Residuos y Recursos Hídricos** y **Legislación, Política y Procedimientos Ambientales**, seguidos por **Información, Educación y Participación Pública**. A partir de 2010, la actividad disminuye de forma significativa, manteniéndose en niveles bajos pero estables hasta 2020, con repuntes visibles en **Gestión de Residuos y Recursos Hídricos**, **Información y Participación Pública** y **Legislación y Política Ambiental** en años como 2018 y 2020.

La categoría de **Gestión de Residuos y Recursos Hídricos** destaca por su continuidad a lo largo del periodo, mostrando presencia constante incluso en los años de menor actividad general. Por su parte, **Legislación y Política Ambiental** tuvo un papel relevante durante la segunda mitad de la década de 2000, pero disminuyó notablemente en los años posteriores. En conjunto, el gráfico sugiere una etapa de intensa actividad institucional y educativa en torno a temas ambientales entre 2006 y 2009, seguida por una reducción progresiva en la visibilidad o producción de contenidos sobre estas temáticas durante la década siguiente, aunque con señales de reactivación hacia el final del periodo analizado.

Los resultados revelan una atención pública oscilante hacia los temas ambientales en México, lo que sugiere una dinámica reactiva ante eventos específicos en lugar de un compromiso sostenido. Esta variabilidad representa un reto para la formulación de políticas ambientales de largo plazo, que requieren continuidad en el interés ciudadano y político. La relativa constancia observada en **Legislación y Política Ambiental** podría aprovecharse como base para fortalecer la implementación normativa y el seguimiento institucional.

Por otro lado, la irregularidad en categorías como **Información, Educación y Participación Pública** pone de manifiesto la necesidad de reforzar las estrategias de sensibilización ambiental, asegurando su continuidad más allá de los contextos de crisis. Fomentar la educación ambiental, especialmente desde las instituciones públicas, puede contribuir a estabilizar el interés en áreas clave como la gestión de residuos, la biodiversidad y la participación ciudadana, en línea con los principios del Acuerdo de Escazú y los compromisos globales en materia de sostenibilidad.

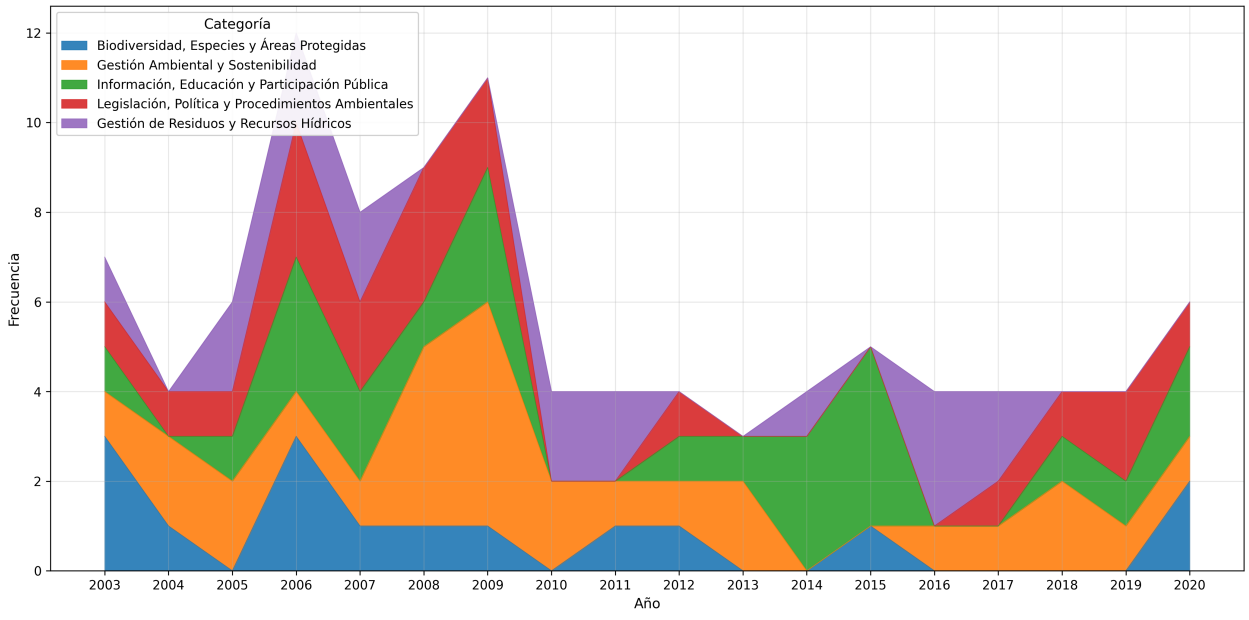

Figura 5.26. Tendencias ambientales a lo largo del tiempo.

5.7 Implicaciones para el acceso a la información en México

En el nuevo entorno normativo del acceso a la información en México, caracterizado por el incremento masivo en el volumen de solicitudes y la necesidad de adoptar enfoques más eficientes de apertura, los instrumentos desarrollados en esta investigación representan una alternativa concreta, escalable y alineada con los principios de la Ley General de Transparencia y Acceso a la Información Pública (LGTAIP, 2025).

La combinación de técnicas de modelado de temas optimizadas mediante algoritmos genéticos ajustando parámetros como α, β y K del modelo LDA junto con la generación automática de descripciones interpretables mediante modelos de lenguaje como GPT, permite automatizar el análisis de millones de solicitudes. Esta metodología identifica patrones temáticos con alta precisión y facilita su comprensión, ofreciendo a las instituciones públicas un mecanismo para anticipar demandas ciudadanas y orientar la apertura de datos con base en necesidades reales y localizadas. Esta propuesta se alinea con el principio de transparencia con sentido social, así como con el mandato de difundir proactivamente la información de interés público (art. 20) y con la implementación de tecnologías orientadas a la publicación automatizada de datos relevantes para la población (art. 54, LGTAIP 2025).

Desde esta lógica, el análisis masivo de solicitudes no se limita a atender requerimientos individuales, sino que posibilita una estrategia de transparencia proactiva territorializada, orientada por el aprovechamiento social de la información pública (Ferencek & Borštnar, 2025).

Del mismo modo, esta propuesta ofrece una base tecnológica accesible para actores no gubernamentales —organizaciones de la sociedad civil, medios de comunicación, universidades— que deseen construir herramientas propias de monitoreo, visualización temática y análisis comparado entre entidades o periodos. La automatización de la clasificación temática y la generación de etiquetas interpretables mediante GPT reduce barreras técnicas y habilita a diversos colectivos ciudadanos para incidir en la formulación de políticas públicas, emitir alertas tempranas o detectar omisiones sistemáticas en la entrega de información (Muhetaer & Hao, 2025).

Además, se abren nuevas posibilidades para repensar las métricas de calidad del acceso a la información. Más allá de los indicadores convencionales como tiempos de respuesta o porcentaje de cumplimiento, esta metodología permite construir indicadores centrados en la coherencia temática, diversidad de tópicos, recurrencia de solicitudes no satisfechas y existencia de vacíos informativos regionales. Estos enfoques se encuentran alineados con el artículo 52 de la LGTAIP, que mandata al Sistema Nacional establecer criterios para evaluar la efectividad de la política de transparencia con sentido social, tomando como base la reutilización y el aprovechamiento que la sociedad haga de la información.

Esta propuesta va más allá de la automatización de procesos: abre nuevas posibilidades para el aprovechamiento estratégico de la información pública. Al facilitar su análisis, comprensión y uso por parte de distintos sectores sociales, contribuye al fortalecimiento de una cultura de transparencia activa, situada y orientada a las demandas reales de la ciudadanía. Este enfoque, además, sienta las bases para un modelo de gobernanza informada por evidencia, en el que las decisiones institucionales se alineen con los intereses temáticos expresados directamente por la ciudadanía a través de la PNT.

En este sentido, los resultados obtenidos permiten afirmar que se cumplieron los objetivos planteados en esta investigación. La metodología desarrollada permitió identificar los temas de mayor interés ciudadano, sus variaciones territoriales y temporales, así como generar mecanismos para su interpretación y visualización. Esto confirma la hipótesis de que el análisis automatizado de solicitudes de información pública puede ser una vía efectiva para fortalecer las estrategias de transparencia, adaptándolas a las necesidades reales de cada región del país.

6. Conclusión

En esta investigación se abordó la automatización de la identificación de temas solicitados por los ciudadanos a la Plataforma Nacional de Transparencia. A través de la propuesta metodológica, se implementó la Ley de Zipf para la optimización del vocabulario y se desarrolló un algoritmo genético para la automatización de los hiperparámetros de α, β y el número de tópicos en el modelo LDA, basándose en la coherencia del modelo. Este enfoque demostró una mejora significativa en la eficiencia comparada con la asignación manual de estos valores, representando un avance importante en la identificación y asignación automática de temas.

En el contexto actual de big data, la automatización y optimización de procesos resulta crucial para gestionar el creciente volumen de información. Esta investigación contribuye a mejorar la eficiencia en el análisis de grandes conjuntos de datos y fortalece la transparencia y la rendición de cuentas en el gobierno. La implementación de la Ley de Zipf permitió una optimización más precisa del vocabulario, mejorando la calidad de los temas identificados al filtrar palabras irrelevantes y sugiriendo un conjunto de palabras clave para la creación de un stop word personalizado. El uso de algoritmos genéticos para ajustar los parámetros de LDA resultó en una mayor coherencia y precisión en la clasificación de los temas, superando métodos tradicionales.

Un avance adicional fue la integración de modelos de lenguaje de gran escala (LLM), lo cual permitió la asignación automática de títulos y descripciones de tópicos, capturando de manera efectiva el tema central a partir de probabilidades clave. Este enfoque metodológico innovador representa una contribución significativa al análisis de las solicitudes de información en la Plataforma Nacional de Transparencia, siendo aplicable a cualquier órgano garante de las 32 entidades federativas. Además, su utilidad se extiende a la generación de reportes y la promoción de la transparencia proactiva, permitiendo la apertura de información recurrente en ciertas regiones geográficas y ofreciendo a académicos, periodistas y ciudadanos una herramienta valiosa para analizar y comprender los temas de interés reflejados en el portal de datos abiertos.

Esta propuesta tiene el potencial de facilitar el acceso a la información pública, permitiendo que tanto ciudadanos como autoridades identifiquen con mayor facilidad los temas de interés en grandes volúmenes de datos. Para los ciudadanos, esto facilita una mayor participación en los asuntos gubernamentales, mientras que para el gobierno, fortalece la relación con la ciudadanía al mejorar la calidad y accesibilidad de la información.

Durante el desarrollo de este trabajo, se enfrentaron desafíos como la presencia de faltas de ortografía o palabras mal escritas en muchas de las solicitudes, lo que complicó el procesamiento automático. Además, la capacidad computacional fue un factor limitante al manejar el volumen total de solicitudes; por ello, se optó por dividir los datos por entidad federativa. Sin embargo, estados como el Estado de México, Ciudad de México y Jalisco presentaron un volumen de solicitudes tan elevado que fue necesario subdividir aún más los datos para su procesamiento.

Se podrían explorar mejoras en la corrección automática de errores ortográficos y en la normalización de textos para perfeccionar la entrada de datos al modelo.Además, la adaptación de esta metodología a otras plataformas de información pública o su aplicación en diferentes contextos temáticos o geográficos podría ser un área de expansión.

Los resultados de esta investigación tienen aplicaciones directas tanto para los gestores de la Plataforma Nacional de Transparencia, al mejorar la eficiencia en el análisis y clasificación de temas, como para los usuarios que buscan información específica en grandes volúmenes de datos. Esto facilita el acceso a datos relevantes y fomenta una mayor interacción entre el gobierno y los ciudadanos, promoviendo una gestión pública más transparente y accesible.

7. Apéndice

Para una revisión detallada de todos los tópicos generados y las tablas empleadas en este estudio, se encuentra disponible un conjunto completo de datos y recursos en la siguiente URL:

https://drive.google.com/file/d/1--gN6gRzdEqeeIaK6ahslRpT0OKE_Ppo/view?usp=sharing

Bibliografía

Abdelrazek, A., Eid, Y., Gawish, E., Medhat, W. & Hassan, A. (2023). Topic modeling algorithms and applications: A survey. *Information Systems*, *112*, 102131.

Ackerman, J. & Sandoval, E. I. (2008). *Leyes de Acceso a la Información en el Mundo (cuarta edición)*. Distrito Federal, México, Instituto Federal de Acceso a la Información Pública (IFAI).

Aguilar-Arévalo, M. & Ramírez-Montaño, N. (2019). Diagnóstico de acceso y uso de la información pública para la exigencia colectiva de derechos. México: Nosotro s por la Democracia, 2019 [citado octubre 2019].

Asamblea Nacional Constituyente de Francia. (1789). *Declaración de los Derechos del Hombre y del Ciudadano: Artículo 11* [Recuperado de https://www.conseil-constitutionnel.fr/sites/default/files/as/root/bank_mm/espagnol/es_ddhc.pdf]. https://www.conseil-constitutionnel.fr/sites/default/files/as/root/bank_mm/espagnol/es_ddhc.pdf

Bagozzi, B. E., Berliner, D. & Almquist, Z. W. (2016). Predicting Government (Non) Responsiveness to Freedom of Information Requests with Supervised Latent Dirichlet Allocation, En *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. https://api.semanticscholar.org/CorpusID:43093099

Bagozzi, B. E., Berliner, D. & Almquist, Z. W. (2019). When does open government shut? Predicting government responses to citizen information requests. *Regulation & Governance*. https://api.semanticscholar.org/CorpusID:211437867

Barrios, C. I. B. (2017). Evolución constitucional de de derecho de acceso a la información. *Investigación Cientıfica*, *11*(1), 12-12.

Bautista-Farías, J. (2015). La nueva Ley General de Transparencia: alcances y retos. *Análisis Plural*, *primer semestre*.

Berliner, D., Bagozzi, B. E. & Palmer-Rubin, B. (2018). What information do citizens want? Evidence from one million information requests in Mexico. *World Development*, *109*, 222-232. https://doi.org/10.1016/j.worlddev.2018.04.016

Berliner, D., Bagozzi, B. E., Palmer-Rubin, B. & Erlich, A. (2020). The Political Logic of Government Disclosure: Evidence from Information Requests in Mexico. *The Journal of Politics*, *83*, 229-245. https://api.semanticscholar.org/CorpusID:211405056

Berliner, D., Palmer-Rubin, B., Reyes, J. E. T., Bagozzi, B. & Erlich, A. (2022). Big data y acceso a la información en México.

Blei, D. M. & Lafferty, J. D. (2007). A correlated topic model of Science. *The Annals of Applied Statistics*, *1*, 17-35. https://api.semanticscholar.org/CorpusID:8872108

Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993-1022.

Calistus, U. C., Onyesolu, M. O., Doris, A. C. & Egwu, C. V. (2024). Exploring Latent Dirichlet Allocation (LDA) in Topic Modeling: Theory, Applications, and Future Directions. *NEWPORT INTERNATIONAL JOURNAL OF ENGINEERING AND PHYSICAL SCIENCES*. https://api.semanticscholar.org/CorpusID:268368446

Cañete, J. & et al. (2020). Spanish pre-trained BERT model and evaluation data. *PML4DC at ICLR 2020*, 1-10.

Carbonetto, P., Sarkar, A., Wang, Z. & Stephens, M. (2021). Non-negative matrix factorization algorithms greatly improve topic model fits. *arXiv preprint arXiv:2105.13440*.

CEPAL. (2018). *Acuerdo Regional sobre el Acceso a la información, la Participación Pública y el Acceso a la Justicia en Asuntos Ambientales en América Latina y el Caribe*. http://www.cepal.org/acuerdodeescazu

Cépeda, G. C. (2018). Sobre la transparencia, el derecho de acceso a la información pública y los deberes de publicidad. Su diferencia conceptual y práctica. *Revista de Derecho Público*, 75-95. https://api.semanticscholar.org/CorpusID:155676609

Colín, A. I. & Huesca, M. S. A. (2022). El derecho de acceso a la información pública y las lenguas indígenas en México. *Desafios Jurídicos*. https://api.semanticscholar.org/CorpusID:252022588

Coria, E. G. C. & López, P. R. F. (2024). Guıa metodológica para el uso de mineria de datos en la Plataforma Nacional de Transparencia. *Estudios en derecho a la información*, *1*(17), 61-75.

Cotino, L. (2013). Del "deber de publicidad" de Brandeis al "open government" de Obama. Regulación y control de la información pública a través de las nuevas tecnologías (G. E. Roca, Ed.). En G. E. Roca (Ed.), *La protección de los derechos humanos por las defensorías del pueblo: Actas del I Congreso Internacional del PRADPI*.

CPEUM. (1917). Constitución Política de los Estados Unidos Mexicanos.

Cueto, G. A. T. (2017). La homogeneización del derecho de acceso a la información en México. La ruptura de la desigualdad en la garantía de este derecho a partir de la Ley General de Transparencia. *Entas*, *1*(931).

Curich, Y. L. N. (2016). El derecho de acceso a la información pública: contenido e importancia. *Forseti. Revista de derecho*, *4*(6), 127-145.

Dang, T. & Nguyen, V. T. (2018). ComModeler: Topic Modeling Using Community Detection., En *EuroVA@ EuroVis*.

Darbishire, H. (2010). *Proactive Transparency: The future of the right to information?* World Bank.

Dieng, A. B., Wang, C., Gao, J. & Paisley, J. W. (2016). TopicRNN: A Recurrent Neural Network with Long-Range Semantic Dependency. *ArXiv*, *abs/1611.01702*. https://api.semanticscholar.org/CorpusID:6039192

DOF. (2015). Ley General de Transparencia y Acceso a la Información Pública.

DOF. (2016). ACUERDO del Consejo Nacional del Sistema Nacional de Transparencia, Acceso a la Información Pública y Protección de Datos Personales, por el que se aprueban los Lineamientos para la implementación y operación de la Plataforma Nacional de Transparencia. *Diario Oficial de la Federación*. https://tinyurl.com/acuerdodof

DUDH. (1948). Declaración Universal de los Derechos humanos. *Naciones Unidas*, *2*.

Ferencek, A. & Borštnar, M. K. (2025). Open Government Data Topic Modeling and Taxonomy Development. *Systems*, *13*(4), 242. https://doi.org/10.3390/systems13040242

Fierro, A. (2021). COVID-19 and the right to access to information. *Revista Juris Poiesis*, *24*(36), 257-266.

Ford, M. O. M. G. (2016). *Plataforma Nacional de Transparencia* (inf. téc.). Instituto Nacional de Transparencia, Acceso a la Información y Protección de Datos Personales.

Fuenmayor E., A. (2004). *El derecho de acceso de los Ciudadanos a la Información Pública. Análisis Jurídico y recomendaciones para una propuesta de ley modelo sobre el derecho de acceso de los ciudadanos a la información pública*. San José, UNESCO San José.

Goldberg, D. E. (1989). Optimization, and machine learning. *Genetic algorithms in Search*.

González, C., Carreón, G., Sánchez, G., Mejía, J. & Hernández, L. (2022). *Informe sobre la situación de las personas y comunidades defensoras de los derechos humanos ambientales en México, 2021*. Ciudad de México, Centro Mexicano de Derecho Ambiental.

Grau, N. C. (2006). La transparencia en la gestión pública: ¿Cómo construirle viabilidad? *Estado, gobierno, gestión pública: Revista Chilena de Administración Pública*, 22-44.

Grootendorst, M. R. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *ArXiv*, *abs/2203.05794*. https://api.semanticscholar.org/CorpusID:247411231

Gutiérrez, E. (2008). *La transparencia*. Nostra Ediciones. https://books.google.com.mx/books?id=OIUVAQAAIAAJ

Hasan, M., Rahman, A., Karim, M. R., Khan, M. S. I. & Islam, M. J. (2021). Normalized approach to find optimal number of topics in Latent Dirichlet Allocation (LDA), En *Proceedings of International Conference on Trends in Computational and Cognitive Engineering: Proceedings of TCCE 2020*. Springer.

Héritier, A. (2003). Composite democracy in Europe: the role of transparency and access to information. *Journal of European public policy*, *10*(5), 814-833.

Hofbauer, H. & Cepeda, J. (2005). Transparencia y rendición de cuentas. En M. Merino (Ed.), *Transparencia: libros, autores e ideas*. México, IFAI-CIDE.

Instituto Nacional de Transparencia, Acceso a la Información y Protección de Datos Personales (INAI). (2023). Base de datos de solicitudes de acceso a la información pública [Datos consultados en 2023]. https://www.plataformadetransparencia.org.mx/

Isensee Rimassa, C. A. & Muñoz Severino, J. (2010). Principio constitucional de transparencia y su materialización en el derecho de acceso a la información pública: análisis crıtico de su regulación legal y administrativa.

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y. & Zhao, L. (2019). Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. *Multimedia tools and applications*, *78*, 15169-15211.

Kingma, D. P., Welling, M. Et al. (2019). An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, *12*(4), 307-392.

Kuri-Morales, A. F., Aldana-Bobadilla, E. & López-Peña, I. (2013). The best genetic algorithm II: A comparative study of structurally different genetic algorithms, En *Mexican International Conference on Artificial Intelligence*. Springer.

Lai, Y.-W. & Chen, M.-Y. (2023). Review of Survey Research in Fuzzy Approach for Text Mining. *IEEE Access*, *11*, 39635-39649. https://doi.org/10.1109/ACCESS.2023.3268165

Landauer, T. K., Foltz, P. W. & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*(2-3), 259-284.

Lee, D. D. & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, *401*, 788-791. https://api.semanticscholar.org/CorpusID:4428232

Levy, O. & Goldberg, Y. (2014). Linguistic regularities in sparse and explicit word representations, En *Proceedings of the eighteenth conference on computational natural language learning*.

Li, H. & Yamanishi, K. (2003). Topic analysis using a finite mixture model. *Information Processing & Management*, *39*(4), 521-541.

López, S. (2009). El acceso a la información como un derecho fundamental: la reforma al artículo 6 de la Constitución mexicana. *Cuadernos de transparencia*, *17*.

Martínez, A. C. & de Mingo, A. C. (2018). El impacto de la gestión documental en la transparencia de las Administraciones públicas: la transparencia por diseño. *Gestión y análisis de polıticas públicas*, 6-16.

Martínez Díaz, M. E., Heras Gomez, L. L. Et al. (2011). Transparencia gubernamental y acceso a la información en México (2002-2010): un analisis exploratorio.

Mayer-Foulkes, D. (2018). Efficient Urbanization for Mexican Development. *International Journal of Economics and Finance*, *10*(10), 1-1.

McCreadie, M. & Rice, R. E. (1999). Trends in analyzing access to information. Part I: cross-disciplinary conceptualizations of access. *Information Processing & Management*, *35*(1), 45-76.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, *26*.

Mitchell, M. (1998). *An introduction to genetic algorithms*. MIT press.

Mohr, J. W. & Bogdanov, P. (2013). Introduction—Topic models: What they are and why they matter. Elsevier.

Moody, C. E. (2016). Mixing Dirichlet Topic Models and Word Embeddings to Make lda2vec. *ArXiv*, *abs/1605.02019*. https://api.semanticscholar.org/CorpusID:15747275

Moreno, M. C. C. & Castro, G. L. G. (2023). Unveiling Public Information in the Metaverse and AI Era: Challenges and Opportunities. *Metaverse Basic and Applied Research*, *2*, 35-35.

Morin, F. & Bengio, Y. (2005). Hierarchical probabilistic neural network language model, En *International workshop on artificial intelligence and statistics*. PMLR.

Muhetaer, M. & Hao, F. (2025). Exploring the Application of ChatGPT in Scientific Topic Analysis: A Novel Paradigm for Enhanced Analysis and Efficiency. *Applied Intelligence*, *55*(7), 625. https://doi.org/10.1007/s10489-025-06498-y

Muñoz, D. (2023). El acuerdo de Escazú en México, a un año de su implementación. *InterNaciones*, *1*(24), 183-208.

Naciones Unidas. (1966). Pacto Internacional de Derechos Civiles y Políticos.

Neri, R. A. O. (2022). Brechas digitales y territorio: los entornos tecnológicos-digitales en las viviendas mexicanas. *Revista Ra Ximhai*, *18*(4 Especial), 103-125.

Newman, D., Lau, J. H., Grieser, K. & Baldwin, T. (2010). Automatic evaluation of topic coherence, En *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

Olivos-Fuentes, M. (2012). *El derecho a la información pública municipal*. Editorial Novum.

Omar, M., On, B.-W., Lee, I. & Choi, G. S. (2015). LDA topics: Representation and evaluation. *Journal of Information Science*, *41*(5), 662-675. https://doi.org/10.1177/016555151558783

ONU. (2014). *A world that counts: Mobilizing the data revolution for sustainable development* (inf. téc.). Independent Expert Advisory Group on a Data Revolution for Sustainable Development.

Organización de los Estados Americanos. (2020). Ley modelo interamericana sobre acceso a la información pública (2.0).

Osollo, A. G. R. (2021). El derecho a la privacidad y la protección de datos personales transfronterizos. *Revista Eurolatinoamericana de Derecho Administrativo*, *8*(1), 35-60.

Padilla, A. R., Storie, J., Storie, C. D. & Herrera, J. M. E. (2023). Urbanization in the Mexico City Metropolitan Area 1900–2020: Urban Dynamics and Driving Factors. *Cartographica: The International Journal for Geographic Information and Geovisualization*, *58*(4), 189-204.

Pathik, N. & Shukla, P. (2020). Simulated annealing based algorithm for tuning LDA hyper parameters, En *Soft Computing: Theories and Applications: Proceedings of SoCTA 2019*, Springer Singapore.

Pérez, C. C. F. & Cecilia, C. S. (2019). La comunicación social en México y la lucha fallida por la transparencia: Del Imperio Azteca a la Ley Chayote, En *Documentos Académicos – UA Docencia Superior*, Universidad Autónoma de Zacatecas. https://doi.org/10.48779/th1q-fh23

Perramon, J. (2013). La transparencia: concepto, evolución y retos actuales. *Revista de contabilidad y dirección*, *16*, 11-27.

Piantadosi, S. T. (2014). Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review*, *21*(5), 1112-1130. https://doi.org/10.3758/s13423-014-0585-6

PIDCP. (2018). Pacto Internacional de Derechos Civiles y Políticos [Accedido el 7 de julio de 2024]. https://api.semanticscholar.org/CorpusID:159771769

Ramos, T. D. (2020). El derecho de acceso a la información pública. La ineficacia de su ejercicio en México. *Ciencia jurıdica*, *9*(18), 21-39.

Rashid, J., Shah, S. M. A. & Irtaza, A. (2019). Fuzzy topic modeling approach for text mining over short text. *Information Processing & Management*, *56*(6), 102060.

Rieger, J., Koppers, L., Jentsch, C. & Rahnenführer, J. (2020). Improving reliability of Latent Dirichlet Allocation by assessing its stability using clustering techniques on replicated runs. *ArXiv [Cs.CL]*. https://doi.org/10.48550/ARXIV.2003.04980

Rijcken, E. (2023). C Topic Coherence Explained: Understanding the metric that correlates the highest with humans. *Towards Data Science*.

Rijcken, E., Scheepers, F., Zervanou, K., Spruit, M., Mosteiro, P. & Kaymak, U. (2023). Towards Interpreting Topic Models with ChatGPT, En *The 20th World Congress of the International Fuzzy Systems Association*.

Rincón, A. C. G. (2020). Acceso a la información y protección de datos en México en tiempos de la pandemia.¿ qué esperar de un gobierno abierto y responsable? *Iuris Tantum*, *34*(31), 45-55.

Röder, M., Both, A. & Hinneburg, A. (2015). Exploring the space of topic coherence measures, En *Proceedings of the eighth ACM international conference on Web search and data mining*. https://doi.org/10.1145/2684822.2685324

Ruiz, F. J. D. (2009). Retos y oportunidades de la administración y el gobierno electrónicos: Derecho a las TIC y alfabetización digital. *Zona Próxima*, *1*(10), 104-125.

Sandoval Ballesteros, I. E. (2008). *Leyes de acceso a la información en el mundo*. Instituto Nacional de Transparencia, Acceso a la Información y Protección de Datos Personales.

Sanz Salguero, F. J. (2016). Relación entre la protección de los datos personales y el derecho de acceso a la información pública dentro del marco del derecho comparado. *Ius et Praxis*, *22*(1), 323-376.

Sbalchiero, S. & Eder, M. (2020). Topic modeling, long texts and the best number of topics. Some Problems and solutions. *Quality & Quantity*, *54*, 1095-1108.

Steinberger, J. & Jezek, K. (2004). Using latent semantic analysis in text summarization and summary evaluation, En *Proc. ISIM*.

Stevens, K., Kegelmeyer, P., Andrzejewski, D. & Buttler, D. (2012). Exploring topic coherence over many models and many topics, En *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*.

Subeno, B. (2017). Optimization Number of Topic Latent Dirichlet Allocation.

Valencia, G. A. (2011). Importancia del derecho de acceso a la información. *La Revista de Derecho*, *32*, 29-47.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Vergara, R. (2007). *La transparencia como problema* (Vol. 5). IFAI.

Villanueva, E. (2008). *Derecho de la información: doctrina, legislacion, jurisprudencia*. Ediciones CIESPAL.

Walby, K. & Larsen, M. (2012). Access to information and freedom of information requests: Neglected means of data production in the social sciences. *Qualitative inquiry*, *18*(1), 31-42.

Witness, G. (2021). Última lınea de defensa. Las industrias que causan la crisis climática y los ataques contra personas defensoras de la tierra y el medioambiente. *Global Witness*, 26.

Yannoukakou, A. & Araka, I. (2014). Access to government information: Right to information and open government data synergy. *Procedia-Social and Behavioral Sciences*, *147*, 332-340.

Yarnguy, T. & Kanarkard, W. (2018). Tuning Latent Dirichlet Allocation parameters using ant colony optimization. *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*, *10*(1-9), 21-24.

Zbigniew, M. (1996). Genetic algorithms+ data structures= evolution programs. *Comput Stat*, 372-373.

Zhao, W., Chen, J. J., Perkins, R., Liu, Z., Ge, W., Ding, Y. & Zou, W. (2015). A heuristic approach to determine an appropriate number of topics in topic modeling, En *BMC bioinformatics*. Springer.