Universidad de La Habana
Facultad de Matemática y Computación

Recuperación Semántica de Información: Un enfoque integrado.

Autor:
Laura Victoria Riera Pérez

Tutores:

Lic. Carlos León González
Dra. C. Lucina García Hernández

Trabajo de Diploma
presentado en opción al título de
Licenciado en Ciencia de la Computación

29 de diciembre de 2023

https://github.com/LRiera24/semantic-information-retrieval

Dedicatoria

Agradecimientos

Agradecimientos

Opinión de los tutores

Opiniones de los tutores

Resumen

Resumen en español

Abstract

Resumen en inglés

Índice general

Índice de figuras

Ejemplos de código

Introducción

A lo largo de los siglos, mentes ilustres como las de Descartes, Newton y Bacon han tejido la noción cautivadora de que *el conocimiento es poder*. En la construcción de este conocimiento, la información desempeña un papel fundamental, siendo el material primario del cual se extraen ideas, conceptos y comprensiones profundas. En la sociedad contemporánea, este papel adquiere una relevancia sin precedentes, consolidándose la información como una fuerza motriz esencial que impulsa los engranajes del progreso y facilita la toma de decisiones cruciales.

El origen de la Recuperación de Información, ese arte y ciencia de extraer conocimiento de vastos conjuntos de datos, estuvo impulsado por la necesidad de superar desafíos en el acceso a información. La gestión manual de registros, especialmente en el ámbito de las publicaciones científicas y los archivos de bibliotecas **manning__introduction__2009**, implicaba una labor intensiva y propensa a errores.

A medida que las computadoras comenzaron a desarrollarse y evolucionar, se reconocieron sus capacidades para manejar grandes volúmenes de datos y facilitar la recuperación de información. El uso de computadoras con este propósito se remonta a mediados del siglo XX. Durante la Segunda Guerra Mundial, Alan Turing y las computadoras británicas Colossus fueron fundamentales para procesar y descifrar mensajes encriptados nazis **hodges__alan__2012**. En los 1950s, se implementaron sistemas como el General Electric, que buscaba más de 30,000 resúmenes de documentos, representando un hito inicial en el uso de computadoras para gestionar grandes conjuntos de información. **sanderson__history__2012** Durante la década de 1960, se destacaron avances en la formalización de algoritmos para clasificar documentos en relación con una consulta. Un enfoque destacado consideraba documentos y consultas como vectores en un espacio N-dimensional **sanderson__history__2012**. Durante los años 1970, se produjeron avances significativos, como la complementación de los pesos de frecuencia de término (tf) de Luhn, basados en la ocurrencia de palabras dentro de un documento, con el trabajo de Spärck Jones sobre la ocurrencia de palabras en el conjunto de documentos de una colección **sanderson__history__2012**. Con el auge de las computadoras personales en los 1980s **magazine** y la irrupción de la World Wide Web en 1991 **sendall__world-wide__nodate**, se transformó radicalmente el panorama, con hitos como Yahoo! en 1994 **freeman__yahoo__2008** y el

algoritmo PageRank de Google en 1996 **sanderson__history__2012**, marcando un nuevo paradigma orientado a la web. En los 2000 se presenció una transición hacia la personalización y la búsqueda semántica, con eventos destacados como el lanzamiento del algoritmo Hummingbird por Google en 2013 **lin__how__2014**.

A partir de 2020, las tendencias fundamentales en la evolución de la Recuperación de Información han impulsado una exploración profunda en áreas como la Inteligencia Artificial, el aprendizaje de máquinas y el procesamiento del lenguaje natural, con el objetivo de perfeccionar la precisión de los resultados de búsqueda. Este período ha sido testigo del surgimiento de enfoques innovadores como la búsqueda conversacional y la generalización a la gran mayoría de aplicaciones de la personalización y recomendación de contenido, mejorando directamente la experiencia del usuario; así como la representación semántica y relacional, integrando la esencial comprensión contextual. Además, se han explorado métodos para la extracción de patrones y relaciones complejas, y la clasificación y organización temática.

La transformación significativa de la Recuperación de Información, ha pasado de ser un ámbito exclusivo de profesionales a involucrar a cientos de millones de personas en la búsqueda diaria de información, provocando un impacto profundo en diversas esferas. En el ámbito académico, los Sistemas de Recuperación de Información (SRI) no solo agilizan la investigación, sino que también fomentan la colaboración entre investigadores y facilitan el acceso a recursos compartidos, promoviendo así la difusión eficiente de conocimientos. Paralelamente, en el entorno empresarial, estos sistemas contribuyen a la productividad simplificando la búsqueda de información esencial y respaldando decisiones estratégicas. Asimismo, en redes sociales, personalizan la experiencia del usuario y fortalecen la conexión con recomendaciones adaptadas, mientras que en la industria médica, agilizan diagnósticos y tratamientos. También, en el ámbito cultural, contribuyen a la preservación y acceso a archivos históricos y museos virtuales.

A pesar de estos avances notables, persisten desafíos significativos en los SRI. La ambigüedad semántica, consultas vagas que dificultan la comprensión precisa de la intención del usuario, junto con problemas de relevancia, como el ruido de información y la sensibilidad al contexto, plantean obstáculos a la eficacia de la recuperación de información. Limitaciones tecnológicas, como la dificultad para manejar información multimedia y desafíos en el procesamiento del lenguaje natural, también presentan retos. Además, las preocupaciones éticas, la privacidad del usuario y la adaptación a nuevos contenidos emergentes son áreas críticas a abordar.

En el colectivo de Sistemas de Información de la Facultad de Matemática y Computación (MATCOM) de la Universidad de La Habana, se evidencia una rica trayectoria de investigación y logros en diversas temáticas. Entre los antecedentes, se encuentran trabajos como el presentado por Quintana Wong, García Hernández, Guillot Jiménez y Amable Ambrós en COMPUMAT 2019, que abordó re-

comendaciones para la promoción de la salud mediante el uso de bases de datos NoSQL **quintana-wong_recommendations_2019**. También, se resalta la investigación de Quintana Wong, García Garrido y García Hernández en IWOR 2019 **quintana-wong_integrating_2019**, donde integraron modelos de vecindario y factores latentes para obtener recomendaciones precisas. La participación en eventos internacionales, como la Escuela Latinoamericana de Verano en Investigación Operativa (ELAVIO) 2019 en Lleida, España, refleja el compromiso del colectivo con trabajos que exploraron la combinación de filtrado colaborativo y modelación de tópicos para la recomendación de información **quintana-wong_latent_2019**. En ese mismo año también se destaca el trabajo de Leon González y García Garrido, el cual abordó modelos de generación de tópicos con word embedding, explorando modelos probabilistas como LDA y presentando el algoritmo lda2vec **gonzalez_modelos_nodate**. Además, Quintana-Wong, García Hernández, y colaboradores han contribuido con artículos en revistas especializadas, como el estudio sobre sistemas de recomendación en soluciones analíticas transaccionales para la atención médica **quintana-wong_recommendations_2** Prado Romero y colaboradores, por su parte, han destacado en la predicción de la popularidad de temas en proveedores de noticias, presentando sus avances en ICOR 2020 y con un artículo aceptado en el "Intelligent Data Analysis Journal" **prado-romero_time-sensitive_2020**. Asimismo, se destaca la colaboración internacional en un proyecto conjunto con profesores de la Universidad de L'Aquila, Italia, para desarrollar un Sistema de Recuperación de Información relacionada con la COVID-19, llevando a cabo esta iniciativa de manera efectiva durante los meses de mayo a julio de 2020.

El presente trabajo se centrará en uno de los enfoques fundamentales de la clasificación y organización temática: los modelos de tópicos. Estos modelos matemáticos poseen la capacidad de descubrir patrones temáticos subyacentes y organizar documentos de manera automática según su tema, proporcionando información sobre las palabras que componen cada uno. Sin embargo, enfrentan desafíos actuales, como la adaptación a contextos cambiantes, ya que dependen de hiperparámetros cruciales, cuya configuración se basa en información extraída del corpus y que impacta directamente en la deducción de las temáticas. Este ajuste no es automático y depende del análisis experto para su determinación. En entornos dinámicos, donde la cantidad y naturaleza de los tópicos pueden cambiar con el tiempo, esta adaptación constante puede representar un problema. Otro desafío es la subjetividad en la interpretación de temas, ya que no proporcionan un nombre específico para los tópicos y requieren la intervención de expertos para su identificación. Presentan también problemas en la gestión efectiva de la polisemia, la cual implica distinguir entre los diversos significados de una palabra para una interpretación precisa del tópico. La mejora de la robustez y la capacidad para manejar la variabilidad del lenguaje son áreas cruciales de investigación para perfeccionar los modelos de tópicos.

En combinación, se trabajará con los ampliamente utilizados motores de búsqueda, los cuales facilitan el acceso a vastas cantidades de información a millones de personas. A pesar de sus beneficios, enfrentan limitaciones notables, especialmente en la precisión de los resultados de búsqueda. La ambigüedad del lenguaje humano hace que la simple búsqueda por palabras clave ya no sea suficiente, al generarse resultados no deseados o excluirse información relevante. La capacidad limitada para comprender el contexto y la intención del usuario, subrayan la necesidad constante de mejoras en las herramientas de búsqueda en la web. La desorganización en la visualización de los resultados es otro desafío, ya que la creciente cantidad de información en línea dificulta la identificación rápida de datos pertinentes. La eficiencia en la consulta del corpus como parte del proceso de recuperación de resultados relevantes se vuelve una limitación palpable, ya que examinar la totalidad de la base de datos puede desencadenar procedimientos considerablemente más lentos. Esta circunstancia resalta la importancia de implementar estrategias más selectivas y ágiles en la búsqueda de información.

Por tanto, el problema científico a abordar es la dificultad de los modelos de tópicos para resolver eficazmente el ajuste automatico de hiperparámetros en entornos dinámicos, así como la limitación para la precisión semántica de los tópicos.

Considerando los desafíos anteriormente mencionados las preguntas científicas a responder en esta tesis son, ¿puede la implementación de mejoras en los modelos de tópicos, específicamente en la adaptación a contextos cambiantes y la gestión de la subjetividad en la interpretación de temas, fortalecer la eficiencia de los Sistemas de Recuperación de Información? ¿Puede la aplicación de estrategias más selectivas y ágiles, junto con una comprensión mejorada del contexto, incrementar la eficiencia de los motores de búsqueda en la era de la información abundante y diversa? ¿Puede la visualización descriptiva o clasificada de los resultados ampliar la experiencia del usuario en la interacción con los sistemas de búsqueda de información?

El objetivo general de este trabajo consiste en concebir, diseñar e implementar una solución computacional, mediante la creación de un motor de búsqueda que automatice los procesos de los modelos de tópicos, contribuyendo a la adaptación a contextos cambiantes, la gestión de la subjetividad en la interpretación de temas, y la optimización de los resultados.

Para alcanzar el cumplimiento del objetivo general, se proponen los siguientes objetivos específicos:

1. Profundizar en el marco teórico y conceptual de los SRI, dando prioridad a la comprensión de modelos de tópicos, motores de búsqueda y ontologías.

2. Realizar un estudio exhaustivo del estado actual en entornos dinámicos y en la literatura académica sobre SRI, identificando tendencias clave.

3. Concebir y diseñar estrategias para potenciar la adaptabilidad de los modelos de tópicos en entornos dinámicos y perfeccionar la interpretación semántica de temas.

4. Implementar y evaluar las estrategias concebidas, integrándolas en SRI para medir su eficiencia y efectividad sinérgica en el contexto de adaptabilidad y precisión semántica.

El contenido restante de esta tesis se organiza en cuatro capítulos, abarcando las distintas etapas que constituyen el desarrollo del trabajo. En el Capítulo 2, "Marco Teórico-Conceptual", se proporciona un análisis detallado del estado actual de la ciencia y tecnología en las áreas relevantes, sirviendo como fundamento esencial para la investigación y los resultados obtenidos. El Capítulo 3, "Concepción y Diseño de la Solución", aborda la caracterización general de la propuesta computacional, la arquitectura del sistema, la estructura del modelo analítico y los procesos vinculados a la precisión semántica. Detalles técnicos de la implementación del sistema se presentan en el Capítulo 4, "Implementación y Experimentación" donde se explora cualitativa y experimentalmente la validez de la solución implementada, aprovechando las herramientas disponibles. En la parte del desenlace, se exponen las conclusiones, destacando los logros clave en relación con los objetivos planteados, así como las recomendaciones que señalan futuras direcciones de investigación. La bibliografía utilizada para respaldar la base científica de la solución propuesta y los anexos complementarios se incluyen para facilitar la exploración de temas relacionados.

Capítulo 1

Marco Teórico - Conceptual

En la esfera contemporánea de la gestión de la información, los sistemas de recuperación de información desempeñan un papel fundamental. Estos sistemas, que han evolucionado significativamente a lo largo de las últimas décadas, son cruciales para el manejo eficiente de la creciente marea de datos digitales. Según Manning, Raghavan y Schütze (2008) en "Introduction to Information Retrieval", estos sistemas no solo facilitan el acceso rápido a información relevante sino que también contribuyen a la organización y estructuración de grandes volúmenes de datos. Esta relevancia se extiende a través de una variedad de campos, desde la bibliotecología y la informática hasta los negocios, donde la capacidad de recuperar y clasificar información de manera efectiva se ha convertido en una herramienta indispensable para la toma de decisiones y el conocimiento estratégico.

Los sistemas de recuperación de información se han diversificado en varias vertientes especializadas, reflejando la complejidad y diversidad de las necesidades en este campo. Un componente esencial en la recuperación de información textual es el Procesamiento del Lenguaje Natural (NLP). Como señalan Jurafsky y Martin en "Speech and Language Processing" (2019), el NLP permite a las máquinas comprender, interpretar y manipular el lenguaje humano, proporcionando una comprensión profunda del significado y el contexto, crucial para superar desafíos como la ambigüedad y la variedad idiomática. Esta habilidad no solo mejora la interpretación del lenguaje a múltiples niveles, sino que también juega un papel clave en la organización y clasificación de grandes volúmenes de datos textuales. El NLP ha revolucionado la búsqueda y recuperación de información, permitiendo análisis y síntesis más precisos y rápidos, esenciales en campos como la investigación académica. Además, ha mejorado la accesibilidad, facilitando interfaces de usuario más naturales e intuitivas, como en sistemas de preguntas y respuestas y asistentes virtuales.

En cuanto a la recuperación web y bibliográfica, como describen Croft, Metzler y Strohman en "Search Engines: Information Retrieval in Practice" (2009) y George

en "The Elements of Library Research" (2008) respectivamente, se enfocan en desafíos especializados dentro del campo. La recuperación web aborda aspectos como el ranking de páginas y la optimización para motores de búsqueda, mientras que la recuperación bibliográfica se centra en la organización y el acceso a la literatura académica y científica.

Por otro lado, la recuperación de información multimedia, como abordan Lew et al. en "Content-Based Multimedia Information Retrieval" (2006), se ocupa de contenido que incluye imágenes, audio y video. También está la recuperación de información geográfica, crucial en aplicaciones como los sistemas de información geográfica (SIG) y los servicios basados en la localización, como se describe en "Geographic Information Retrieval" de Purves et al. (2018).

En el presente capítulo, se discuten brevemente componentes clave dentro del campo de la recuperación textual que constituyen la base teórica de esta investigación: los modelos de tópicos, ontologías y *embeddings*, así como los antecedentes relevantes para la automatización de la estimación del número de tópicos en un corpus y la asignación de nombres a los mismos.

1.1. Modelos de Tópicos

Los modelos de tópicos, fundamentales en NLP y la recuperación de información, se destacan por su capacidad para explorar y organizar grandes conjuntos de datos textuales. Estos modelos operan identificando tópicos subyacentes en colecciones de documentos, extrayendo patrones significativos en el uso de palabras y revelando así las estructuras temáticas latentes. Esta funcionalidad los convierte en herramientas esenciales para comprender, categorizar y sintetizar información en grandes volúmenes de texto, lo que es crucial en una era donde la cantidad de datos disponibles crece exponencialmente.

El desarrollo histórico de los modelos de tópicos es un viaje fascinante a través de la evolución del NLP. Comenzando con el Análisis Semántico Latente (LSA) en 1990 por Deerwester et al. ("Indexing by latent semantic analysis"), un método pionero que utilizaba la descomposición en valores singulares para identificar estructuras semánticas en grandes colecciones de texto. Esta técnica fue fundamental para sentar las bases de los modelos de tópicos. El Análisis Semántico Latente Probabilístico (pLSA), propuesto por Hofmann en 1999, representó un avance significativo, introduciendo un enfoque probabilístico para modelar la relación entre documentos y tópicos. Sin embargo, pLSA tenía limitaciones, especialmente en la generalización a documentos no vistos durante el entrenamiento.

El gran avance llegó con la introducción de la Asignación Latente de Dirichlet (LDA) por Blei et al. en 2003 ("Latent Dirichlet Allocation"), considerando cada documento como una mezcla de tópicos latentes, donde cada tópico está definido

por una distribución sobre las palabras. Formalmente, para cada documento d, LDA asume una distribución de tópicos θ_d que se extrae de una distribución a priori de Dirichlet. Para cada palabra en el documento, se elige un tópico z de la distribución de tópicos θ_d. Luego, se selecciona una palabra de una distribución de palabras asociada a ese tópico específico. Este proceso se repite a lo largo de todos los documentos y palabras, iterando para ajustar las distribuciones de tópicos y palabras hasta que el modelo refleje adecuadamente la estructura latente de tópicos en los documentos. Este modelo generativo probabilístico ofreció una mayor flexibilidad y capacidad de interpretación, convirtiéndose en el estándar de oro para el modelado de tópicos.

Desde entonces, los modelos de tópicos han continuado evolucionando, integrando enfoques más sofisticados como el Modelo de Tópicos Correlacionados (CTM) y el Modelo de Tópicos Dinámicos (DTM). El CTM, al incorporar correlaciones entre tópicos, ofrece una representación más matizada y realista de cómo se distribuyen los tópicos en documentos, mientras que el DTM introduce una perspectiva temporal, analizando cómo los tópicos evolucionan con el tiempo. Además, los modelos jerárquicos como el Hierarchical Latent Dirichlet Allocation (HLDA) han proporcionado una estructura más compleja, permitiendo la detección de tópicos en distintos niveles de granularidad. Estos avances han enriquecido el análisis de tópicos, ofreciendo una visión más profunda y detallada de las estructuras temáticas en grandes volúmenes de texto.

1.2. Ontologías

Las ontologías, en el contexto de la informática y NLP, son estructuras de datos que representan conocimientos de manera organizada y jerárquica. Según Guarino (1998) en "Formal Ontologies and Information Systems", una ontología define un conjunto de conceptos y categorías que representan un dominio de conocimiento, así como las relaciones entre estos conceptos. Computacionalmente, las ontologías se manejan frecuentemente como grafos, donde los nodos representan conceptos o entidades, y las aristas o bordes representan las relaciones entre ellos. Las ontologías permiten que las máquinas "comprendan" y procesen el significado de la información de manera más eficaz, fundamental para la extracción y clasificación de información, donde se requiere no solo identificar datos, sino también comprender su semántica.

Según Uschold y Grüninger (1996), las ontologías son creadas principalmente por expertos en un dominio específico, con el objetivo de facilitar la comunicación y la comprensión común en diferentes campos. Estos expertos utilizan las ontologías para establecer un marco conceptual compartido, lo que ayuda a superar las barreras de comunicación y a mejorar la interoperabilidad entre sistemas y organizaciones, lo cual es fundamental en áreas como la ingeniería de sistemas, la integración empresarial y el desarrollo de software.

Las ontologías tienen aplicaciones diversas en la informática y el NLP. Berners-Lee et al. (2001) resaltan su uso en la Web Semántica, al mejorar la accesibilidad y gestión de información, facilitando búsquedas más eficientes y una navegación intuitiva. Hotho et al. (2002) y Lastra-Díaz et al. (2019) exploran su uso en la agrupación de documentos y la similitud semántica. Estas técnicas mejoran la precisión en la categorización y búsqueda de documentos, permitiendo a los sistemas informáticos identificar conexiones y similitudes en el contenido textual basándose en significados subyacentes. Batet et al. (2009) y Corcho (2006) se centran en la clasificación y anotación de documentos, demostrando cómo las ontologías enriquecen la recuperación y gestión de información. Su integración proporciona una mayor profundidad en el análisis de textos, permitiendo la creación de metadatos más ricos y relevantes, lo que mejora sustancialmente la recuperación y gestión de información en bases de datos y repositorios digitales.

Pueden ser clasificadas como ontologías de dominio específico y ontologías generales. Las ontologías de dominio específico se enfocan en áreas de conocimiento particulares, brindando un marco detallado para las entidades y relaciones dentro de ese ámbito específico. Por ejemplo, en el campo de la medicina, la "Ontología de la Genómica del Cáncer" (OGC) es una ontología de dominio específico que detalla la terminología, las relaciones y los procesos relacionados con el cáncer y la genómica. Este tipo de ontología es indispensable en aplicaciones médicas y de investigación, proporcionando una estructura precisa para la gestión y el análisis de datos complejos relacionados con enfermedades y tratamientos.

En contraste, las ontologías generales abarcan un conocimiento más amplio, estableciendo un marco general para clasificar y relacionar conceptos de varios dominios. Son esenciales en aplicaciones que demandan un entendimiento generalizado del conocimiento humano. WordNet es un ejemplo significativo en esta categoría; es una base de datos léxica en inglés que organiza palabras en conjuntos de sinónimos (synsets), definiendo relaciones como sinonimia, antonimia y jerarquías de hipónimos (especificaciones) e hiperónimos (generalizaciones).

1.3. Vectores con contenido semántico: *embeddings*

Los *embeddings*, en el contexto de NLP, son representaciones vectoriales de palabras o frases. Almeida y Xexéo (2023) en su estudio "Word Embeddings: A Survey", explican que estos *embeddings* son vectores densos y distribuidos de longitud fija, construidos utilizando estadísticas de co-ocurrencia de palabras. Estas representaciones codifican información sintáctica y semántica, transformando el texto en una forma que es manejable para los algoritmos de aprendizaje automático, facilitando tareas como clasificación de texto, análisis de sentimientos y traducción automática.

Existen varios tipos de *embeddings*, cada uno con características únicas. Los *word*

embeddings, con modelos preentrenados como Word2Vec y GloVe, representan palabras individuales como vectores en un espacio multidimensional, capturando su significado y relaciones sintácticas. Los *sentence embeddings*, por otro lado, representan oraciones enteras, permitiendo capturar el contexto más amplio de la oración. Los *contextual embeddings* como los generados por modelos como BERT, van un paso más allá, representando palabras en el contexto de frases o párrafos, lo que permite una comprensión más matizada del significado, especialmente para palabras con múltiples interpretaciones.

Existen distintas investigaciones que ilustran la versatilidad de los *embeddings* en diversas áreas para el análisis de textos y relaciones semánticas.. El estudio de Fang et al. (2016) demuestra la aplicación de *embeddings* en la evaluación de la coherencia de temas en datos de Twitter. Utilizando *embeddings*, se pueden medir con precisión la coherencia y relevancia de los tópicos generados, lo que es crucial para la interpretación efectiva de grandes conjuntos de datos sociales. Además, Kuzi et al. (2016) exploran cómo los *embeddings* pueden mejorar la expansión de consultas en motores de búsqueda, proporcionando una búsqueda más rica y contextualizada. Otros ejemplos incluyen el trabajo de Kusner et al., que examina cómo los *embeddings* de palabras se utilizan para medir distancias entre documentos, y Lezama-Sánchez et al. (2022), que investiga el uso de *embeddings* basados en relaciones semánticas para mejorar el análisis de texto. Además, Liu et al. exploran los *embeddings* temáticos, mostrando su aplicación en la comprensión de tópicos específicos en grandes conjuntos de datos.

1.4. Estado del arte

La sección del estado del arte explora desarrollos recientes y metodologías clave en los dos problemas centrales a abordar en este trabajo: la automatización de la identificación del número de tópicos en un corpus y de la asignación de nombres a tópicos.

1.4.1. Identificación del número de tópicos presentes en un corpus

La identificación del número de tópicos en un corpus es crucial en el análisis de datos y el NLP, ya que define la estructura y claridad en la interpretación de grandes volúmenes de texto. Hasta ahora, este proceso depende en gran medida de la intervención de expertos, lo que conlleva una subjetividad inherente y limita la escalabilidad y consistencia en el análisis. La ausencia de métodos automáticos para determinar el número óptimo de tópicos representa un desafío significativo, ya que un enfoque automatizado y objetivo podría adaptarse mejor a la naturaleza dinámica y

variable de los datos, permitiendo a los analistas centrarse más en la interpretación y aplicación de los resultados.

El enfoque de Arun et al. (2010) en "On Finding the Natural Number of Topics with Latent Dirichlet Allocation" se centra en determinar el número óptimo de tópicos para modelos LDA. Proponen una medida basada en la divergencia simétrica de Kullback-Leibler de las distribuciones salientes de los factores de la matriz LDA. Esta medida identifica el número de tópicos observando un 'pico' en los valores de divergencia para el número correcto de tópicos. El método es validado con conjuntos de datos reales y sintéticos.

Por otro lado, el artículo de Gan y Qi (2021) en "Selection of the Optimal Number of Topics for LDA" se centra en una metodología integral para determinar el número óptimo de tópicos en LDA. Presentan un índice que evalúa varios factores: perplejidad, aislamiento de tópicos, estabilidad y coincidencia. Este índice tiene como objetivo lograr una alta capacidad predictiva, un buen aislamiento entre tópicos, evitar tópicos duplicados y asegurar la repetibilidad. Se validó con datasets generales y una aplicación específica en la clasificación de políticas de patentes en China, mostrando buenos resultados.

Vangara et al. (2021) en "Finding the Number of Latent Topics With Semantic Non-Negative Matrix Factorization" introduce SeNMFk, una metodología que combina la factorización de matrices no negativas (NMF) con información semántica para determinar el número óptimo de tópicos. SeNMFk utiliza la divergencia de Kullback-Leibler y se enfoca en la estabilidad de los tópicos a través de un ensamble aleatorio de matrices. Además, presentan el software pyDNMFk para facilitar la estimación del número de tópicos.

Los siguientes artículos, aunque no persiguen el objetivo de identificar automáticamente el número de tópicos presentes en el corpus, son relevantes ya que emplean razonamientos básicos, análogos a los utilizados en esta investigación. El enfoque de Jiang et al. (2011) en "A Fuzzy Self-Constructing Feature Clustering Algorithm for Text Classification" se basa en un algoritmo de agrupación de características autoconstruido difuso. Este método se enfoca en identificar estructuras latentes en datos de texto, utilizando técnicas de agrupación difusa para manejar la incertidumbre y la ambigüedad inherentes a los datos textuales. Permite una clasificación más flexible y adaptativa de los datos, lo que puede ser especialmente útil en aplicaciones donde las categorías no son claramente definidas o donde los datos pueden pertenecer a múltiples categorías simultáneamente, sin tener que especificar el número de grupos como parámetro de entrada. Por otro lado, el enfoque de Thompson y Mimno (2020) en "Topic Modeling with Contextualized Word Representation Clusters" investiga el uso de agrupaciones de representaciones de palabras contextualizadas, como las de BERT y GPT-2, para el modelado de tópicos. Su metodología se basa en la hipótesis de que estas representaciones contextualizadas pueden capturar polisemia y proporcio-

nar información sintáctica más rica, lo que resulta en una organización de documentos similar a la obtenida con modelos LDA tradicionales.

1.4.2. Asignación de nombres a grupos

La asignación automática de nombres a tópicos en un corpus es de gran importancia pues facilita la comprensión e interpretación de los resultados generados por modelos de tópicos. Actualmente, esta tarea también se realiza principalmente a través de la intervención de expertos, un proceso que puede ser subjetivo y laborioso. La carencia de métodos automáticos para esta asignación representa un desafío significativo en la eficiencia y objetividad del análisis de grandes conjuntos de datos. La automatización de este proceso permitiría una identificación de tópicos más coherente y objetiva, reduciendo la carga de trabajo manual y aumentando la capacidad para manejar abundante información.

El documento "An Ontology-based Semantic Tagger for IE system" de Boufaden (2003), detalla un sistema para etiquetar semánticamente conversaciones en el ámbito de búsqueda y rescate marítimo, y facilitar el proceso de la extracción de información. La metodología combina dos fuentes de conocimiento: una ontología SAR y el diccionario-tesauro Wordsmyth. El proceso se divide en cuatro etapas: extracción de palabras candidatas, anotación semántica, filtrado contextual, y utilización de las palabras anotadas para la resolución de correferencias y el llenado de plantillas de extracción de información. Se ejemplifica con un diálogo donde se etiquetan palabras clave con etiquetas específicas del dominio, como STATUS MISSING-VESSEL y LOCATION.

En "Ontology based Web Page Topic Identification" de Singh Rathore y Roy se describe un método para identificar tópicos en páginas web usando una ontología de dominio específico desarrollada manualmente con este propósito. El proceso inicia con la extracción de palabras clave de tags HTML y la detección de co-ocurrencias de palabras en el texto. Primero, se mapean las palabras clave extraídas de la página web a conceptos en la ontología, utilizando la Distancia de Levenshtein para evaluar su similitud. Las palabras clave se clasifican por relevancia, y se establece un umbral para determinar la adecuación del mapeo. En la primera fase, se consideran las palabras clave más relevantes, y si su correspondencia con un nodo de la ontología es significativa, se sugiere que la página pertenece a ese tema. Si no se alcanza este umbral, la segunda fase incluye todas las palabras clave. Si la combinación de todas las palabras clave mapeadas supera el umbral para un nodo, entonces se asigna ese tema al documento.

El trabajo de Saqlain et al. (2016) sigue un enfoque que utiliza WordNet y TF-IDF para asignar nombres automáticamente a tópicos, combinando técnicas semánticas y estadísticas. El proceso inicia con una agrupación jerárquica de los textos y su

posterior procesamiento, incluyendo la eliminación de palabras comunes y la estandarización de términos. Posteriormente, se identifican términos clave en cada grupo utilizando el cálculo de Term Frequency-Inverse Document Frequency (TF-IDF). Estos términos son procesados a través de WordNet para generar sus hiperónimos, y la frecuencia de estos hiperónimos se calcula dentro del grupo. Los hiperónimos que aparecen con mayor frecuencia se seleccionan como etiquetas para los grupos.

Desambiguación del sentido de las palabras

La desambiguación de sentidos de palabras (WSD) es un campo de la lingüística computacional y el NLP que se enfoca en asignar significados precisos a palabras en contextos específicos, diferenciando entre múltiples significados o sentidos. Esto se hace particularmente necesario al utilizar ontologías ya que una palabra puede corresponder a múltiples conceptos en la ontología, cada uno con un significado distinto de acuerdo al contexto.

El algoritmo de Lesk (1986), es una técnica clásica para WSD. El algoritmo opera comparando las definiciones de una palabra objetivo con las palabras presentes en su contexto inmediato. La definición que tenga el mayor solapamiento es elegida como la más probable. A pesar de su simplicidad, el algoritmo de Lesk ha sido fundamental en el desarrollo de técnicas más avanzadas de WSD, y ha sido implementado en discímiles bibliotecas de NLP. Tiene limitaciones evidentes, como su dependencia de coincidencias exactas de palabras y la posibilidad de que el contexto no ofrezca suficiente información para un solapamiento significativo, lo que puede resultar en la elección de sentidos incorrectos para las palabras. Desde entonces han surgido modificaciones de este algoritmo para añadir semántica al proceso.

Ambos, Edmonds y Agirre (2008), y Bevilacqua et al. (2021), proporcionan un análisis de los algoritmos y aplicaciones en WSD. Los algoritmos basados en grafos son fundamentales en los enfoques basados en el conocimiento, aprovechando estructuras de grafos de recursos léxicos como WordNet y BabelNet. Estos algoritmos ayudan a establecer conexiones entre diferentes sentidos de palabras basándose en sus relaciones semánticas. Por otro lado, en los enfoques supervisados, los modelos neuronales, especialmente aquellos que utilizan arquitecturas de Transformer preentrenadas, han demostrado ser altamente efectivos. Estos modelos aprenden a asociar palabras en contextos específicos con sus sentidos correspondientes, utilizando grandes conjuntos de datos anotados. Además, las técnicas que incorporan glosas o definiciones textuales de inventarios de sentidos, como SensEmBERT y ARES, también han mostrado un rendimiento sobresaliente en WSD. Estos métodos aprovechan las representaciones contextualizadas de palabras y los embeddings de sentido para mejorar la precisión de la desambiguación.

También se destaca el uso de inteligencia artificial y metaheurísticas para WSD.

El artículo de AL-Saiagh et al. (2018) introduce un enfoque híbrido que combina la optimización de enjambres de partículas (PSO) con el recocido simulado, mientras que "A Self-adaptive Genetic Algorithm for the Word Sense Disambiguation Problem" de Wojdan Alsaeedan y Mohamed El Bachir Menai (2015) explora el uso de un algoritmo genético autoadaptativo que ajusta automáticamente las probabilidades de cruce y mutación optimizando el proceso.

Capítulo 2

Concepción y diseño de la solución

Este capítulo explora la arquitectura esencial del programa, centrando su atención en el trayecto desde el Pre-procesamiento Semántico hasta la Recuperación de Información por Tópicos. La Figura 2.1 actúa como guía visual para estas dos etapas.

Se inicia con un preprocesamiento léxico básico, estructurando el texto para las operaciones subsiguientes. La Identificación de la Cantidad de Tópicos adopta un enfoque de agrupación semántica y embeddings, eliminando la necesidad de intervención de expertos para obtener información sobre la distribución del corpus. Esta fase se entrelaza con el modelo de Latent Dirichlet Allocation (LDA) para desentrañar patrones temáticos. La transición hacia la Asignación de Nombres a Tópicos conecta el descubrimiento de tópicos con una ontología de dominio general, seleccionando palabras clave basadas en las probabilidades de LDA. Esta estrategia busca facilitar una asignación contextualizada de nombres.

En pos de aplicación y visualización, se implementa un motor de búsqueda que, al recibir una consulta, devuelve resultados organizados según su relevancia y tópicos (nombrados) asociados, permitiendo una exploración más interpretable.

A través de este recorrido, la arquitectura del programa se presenta como una combinación de componentes interconectados, cada uno contribuyendo a la extracción y comprensión de información significativa de conjuntos de datos textuales.

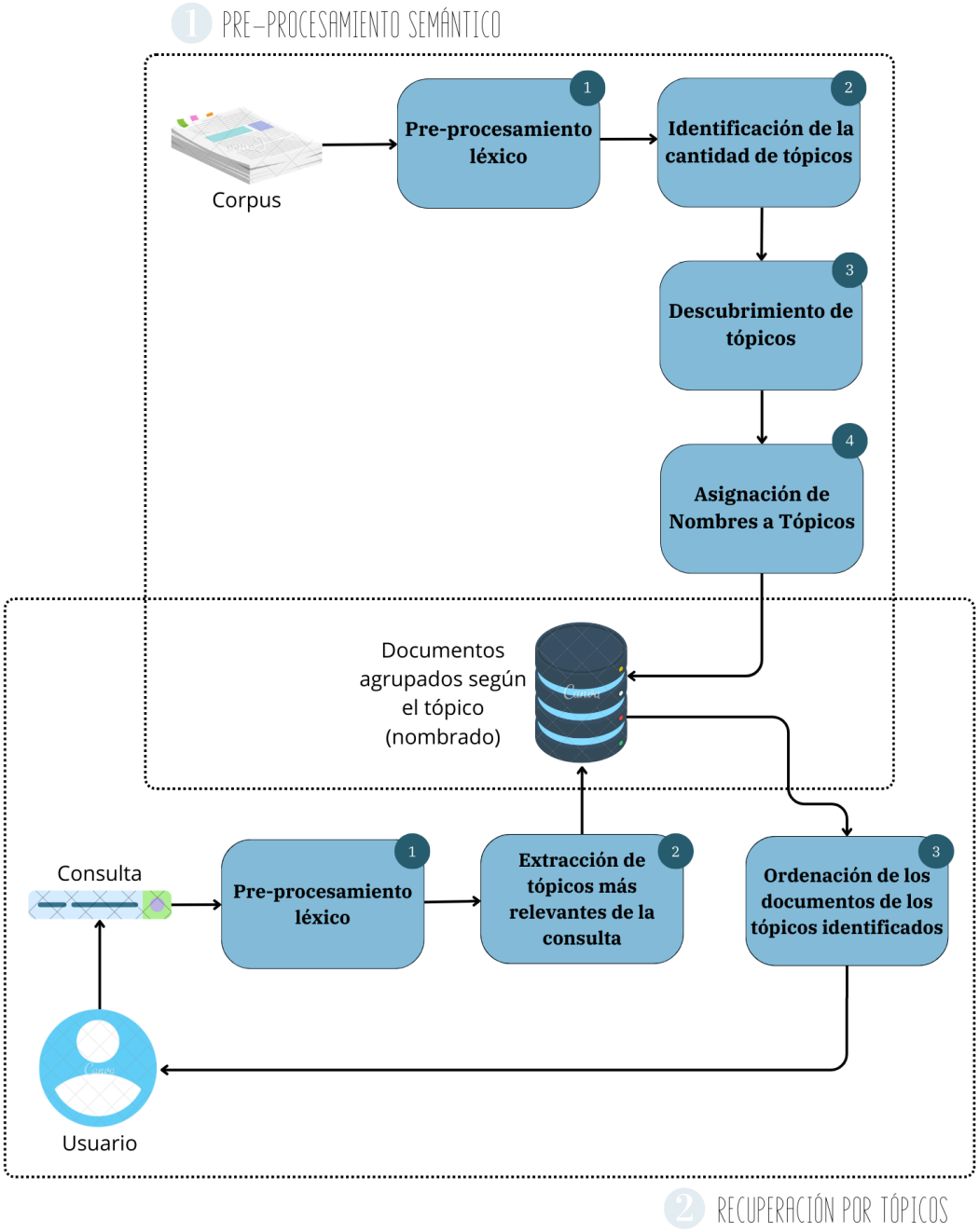

Figura 2.1: Arquitectura del sistema

2.1. Pre-procesamiento semántico

En la búsqueda constante por perfeccionar los modelos de tópicos en el procesamiento de texto, esta sección se enfoca en dos aspectos cruciales: la estimación automática del número de tópicos antes de la aplicación del modelo y la asignación automática de nombres a los tópicos identificados. Estos elementos desempeñan un papel fundamental en la mejora de la interpretación y utilidad de los resultados obtenidos mediante técnicas como Latent Dirichlet Allocation (LDA). La automatización de estos procesos no solo simplifica la implementación de modelos, sino que también enriquece la comprensión de patrones temáticos en grandes conjuntos de datos textuales, allanando el camino hacia un análisis más eficiente y accesible.

2.1.1. Pre-procesamiento léxico

En el ámbito del procesamiento de texto, el preprocesamiento léxico desempeña un papel fundamental al garantizar la adecuada preparación del texto antes de su análisis. Este proceso es esencial en cualquier sistema de procesamiento de texto, ya que establece las bases para la comprensión y extracción de información significativa. En este contexto, se sigue un preprocesamiento léxico básico (ver figura 2.2), un conjunto de pasos iniciales que buscan homogeneizar y organizar el texto de manera que facilite las tareas posteriores de análisis y procesamiento.

La primera etapa del preprocesamiento léxico comienza con la tokenización, fragmentando el texto en unidades léxicas para facilitar la identificación y manipulación de palabras. Seguidamente, se lleva a cabo la eliminación de ruido, suprimiendo elementos redundantes como signos de puntuación o números, que podrían interferir con la interpretación precisa del contenido. A continuación, se realiza la exclusión de términos comunes que carecen de relevancia semántica significativa.

Para la reducción morfológica del vocabulario existen dos técnicas: lemmatización y stemming. La lematización simplifica la identificación de términos al reducir las palabras a sus formas base, mientras que el stemming halla formas truncadas al eliminar sufijos y prefijos. Se escoge la lematización debido al requisito de trabajar con palabras reales para modelos y ontologías en nuestro contexto específico, cosa no garantizada en el stemming. Posteriormente, se aplica un filtrado según la ocurrencia, eliminando términos poco frecuentes o excesivamente repetitivos.

Se construye el vocabulario del corpus y se representan vectorialmente los documentos, eligiendo en este estudio específicamente la representación de bolsa de palabras. Además, se elabora la matriz de co-ocurrencia de términos para capturar las correlaciones entre palabras.

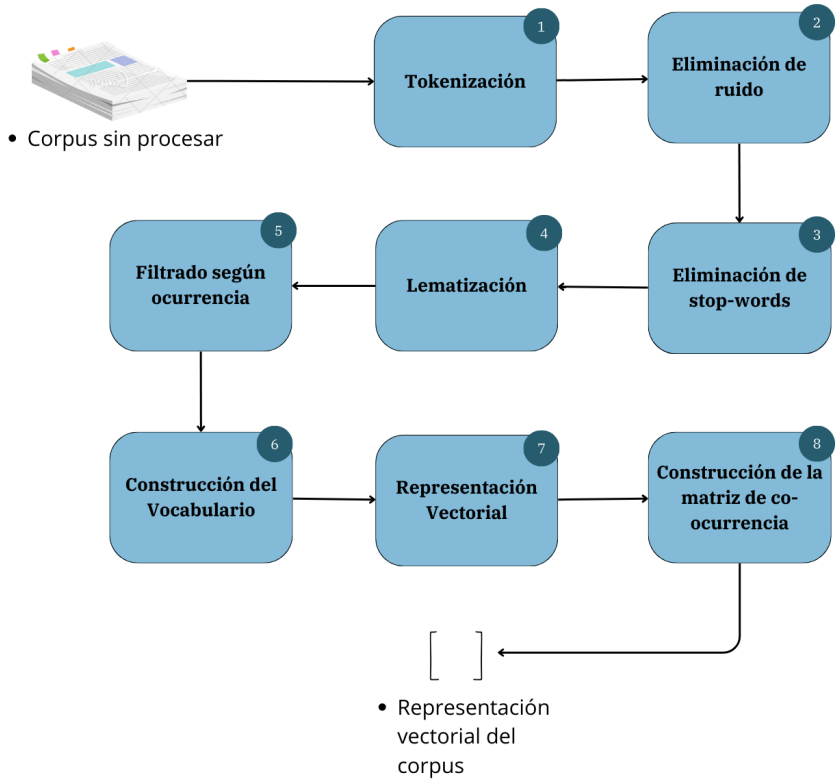

Figura 2.2: Pre-procesamiento léxico

2.1.2. Identificación de la Cantidad de Tópicos

Con el objetivo de identificar la cantidad de tópicos presentes en el corpus se propone un enfoque basado en la agrupación semántica de palabras, representadas por sus embeddings (ver figura 2.3). La hipótesis es que la cantidad de grupos formados, cada uno con una cantidad suficiente de palabras, proporcionará un indicador de la cantidad de tópicos presentes en el corpus. El algoritmo propuesto se basa en que las palabras que comparten significados similares o co-ocurren con frecuencia en documentos estarán asociadas en conjuntos semánticos, revelando así la presencia de tópicos específicos. Al realizar una agrupación flexible en el modelo, donde cada palabra puede pertenecer a varios grupos, se refleja la polisemia y la complejidad semántica del lenguaje.

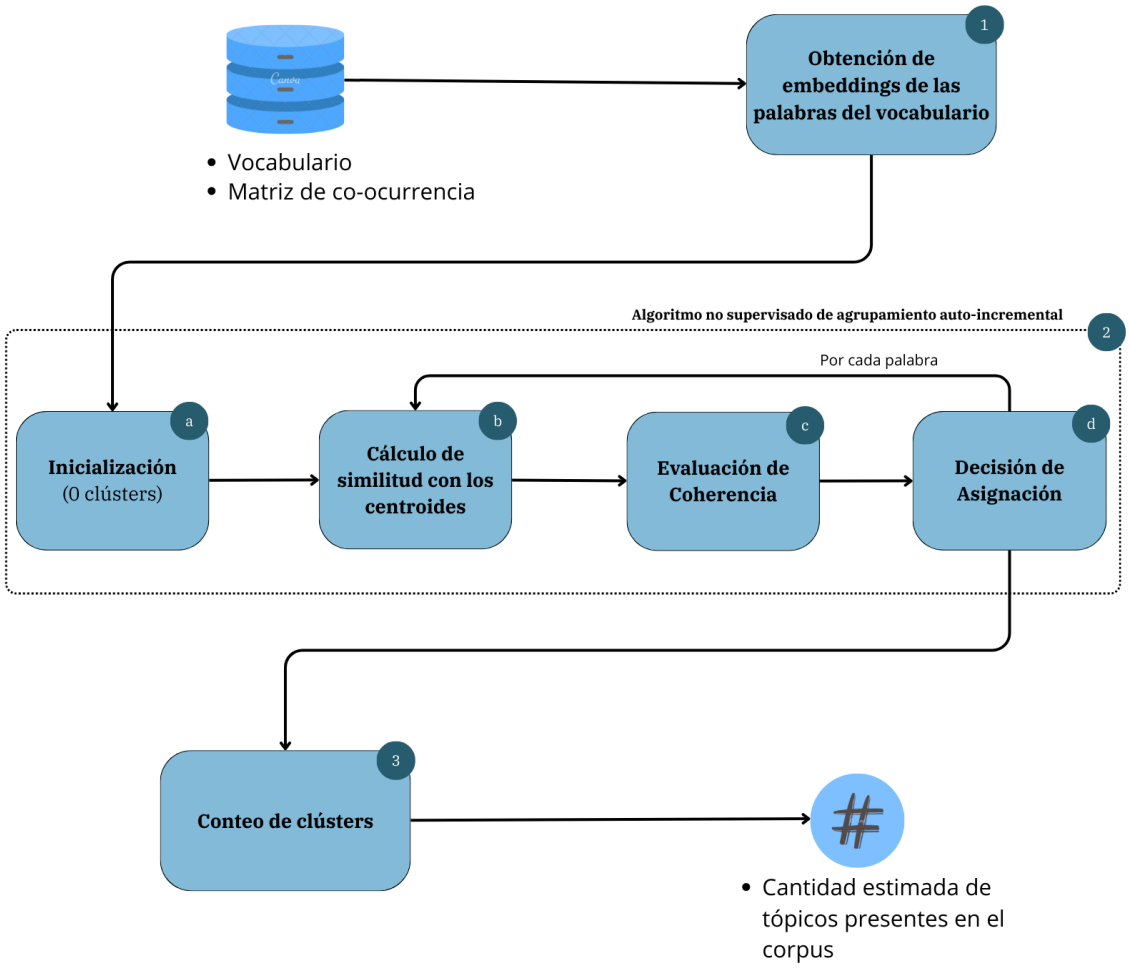

Figura 2.3: Identificación de la cantidad de tópicos

En la fase inicial del proceso, se obtienen los embeddings para las palabras que conforman el vocabulario. Para llevar a cabo este propósito, se emplea un modelo preentrenado con extensas cantidades de texto conocido como Google Word2Vec. Este modelo asigna a cada palabra un vector numérico en un espacio multidimensional. Este espacio está diseñado de tal manera que palabras con significados similares tienen representaciones cercanas. De este modo, cada palabra en el vocabulario se vincula a su correspondiente vector semántico, el cual se utilizará posteriormente en el análisis consecuente. Se escogió este específicamente que es un modelo de dominio general lo cual contribuye a la adaptabilidad del sistema a diferentes contextos.

En la segunda etapa del proceso de identificación de tópicos, se implementa un algoritmo de agrupamiento autoincremental sobre los embeddings de palabras obte-

nidos. A diferencia de los métodos convencionales que requieren una predefinición del número de grupos, este enfoque comienza con 0 grupos y estos se construyen dinámicamente a medida que es necesario.

La similitud semántica, basada en la proximidad en el espacio vectorial, desempeña un papel esencial en la creación de grupos al agrupar palabras con significados similares. Sin embargo, los embeddings, debido a su entrenamiento centrado en co-ocurrencias locales, carecen de información contextual más amplia y no logran distinguir entre las diferentes acepciones de una palabra en distintos contextos. Esta limitación puede resultar en la agrupación errónea de palabras con significado compartido pero usos contextuales diferentes. Para superar este desafío, el enfoque propuesto incorpora la matriz de co-ocurrencia del corpus, enriqueciendo la representación al capturar relaciones contextuales entre palabras. Esta integración mejora la precisión y significado en la formación de grupos.

El algoritmo evalúa la similitud entre el embedding a agregar y los centroides de los grupos, seleccionando aquellos cuya similitud sea igual o superior al umbral especificado. Si el vector no cumple con esta medida para ningún grupo existente, se crea uno nuevo. En caso contrario, se analiza la coherencia contextual con las palabras de cada grupo seleccionado, incorporándolo a aquellos en los que al menos la mitad de las palabras sean coherentes con el vector. Este proceso garantiza que las palabras con significados y contextos afines se agrupen de manera coherente.

En la etapa final, se cuentan los grupos con una cardinalidad superior a un umbral predefinido, que representa la cantidad mínima de palabras necesarias para que un grupo sea considerado como un tema. Esta contabilización proporciona una estimación de la cantidad de tópicos presentes en el corpus.

2.1.3. Descubrimiento de tópicos

La aplicación de modelos de tópicos desempeña un papel fundamental en el descubrimiento de patrones temáticos en grandes conjuntos de datos textuales. Estos modelos proveen una herramienta eficaz para organizar documentos relacionados y revelar las estructuras temáticas subyacentes en un corpus, facilitando así la extracción de información significativa.

Dentro de este contexto, destaca el algoritmo Latent Dirichlet Allocation (LDA), el asigna palabras a tópicos en documentos, utilizando un proceso probabilístico iterativo. Esto facilita la organización y análisis de grandes conjuntos de datos de texto al proporcionar distribuciones que describen la probabilidad de pertenencia de un documento a un tópico y la asociación de una palabra a un tópico específico. A pesar de la existencia de enfoques más complejos, la robustez, aplicabilidad general y estatus clásico de LDA, así como su amplia adopción en la literatura especializada, respaldan su confiabilidad. Esto lo posiciona como una elección sólida en el continuo desarrollo

y mejora de modelos de tópicos para el descubrimiento temático.

2.1.4. Asignación de nombres a tópicos

En el ámbito del análisis de tópicos, conferir nombres de manera automática a los tópicos identificados no solo añade claridad interpretativa, sino que también facilita la comprensión y exploración de grandes conjuntos de documentos. Este proceso, esencial para dotar de significado a los patrones temáticos descubiertos, se llevará a cabo mediante la utilización de una ontología de dominio general. Al aprovechar la riqueza semántica y la estructura jerárquica de esta ontología, se busca lograr una asignación de nombres precisa y contextualizada para cada tópico identificado. Este enfoque (ver figura 2.4) contribuirá a mejorar la interpretabilidad y utilidad de los resultados obtenidos en la fase de descubrimiento de tópicos mediante Latent Dirichlet Allocation (LDA).

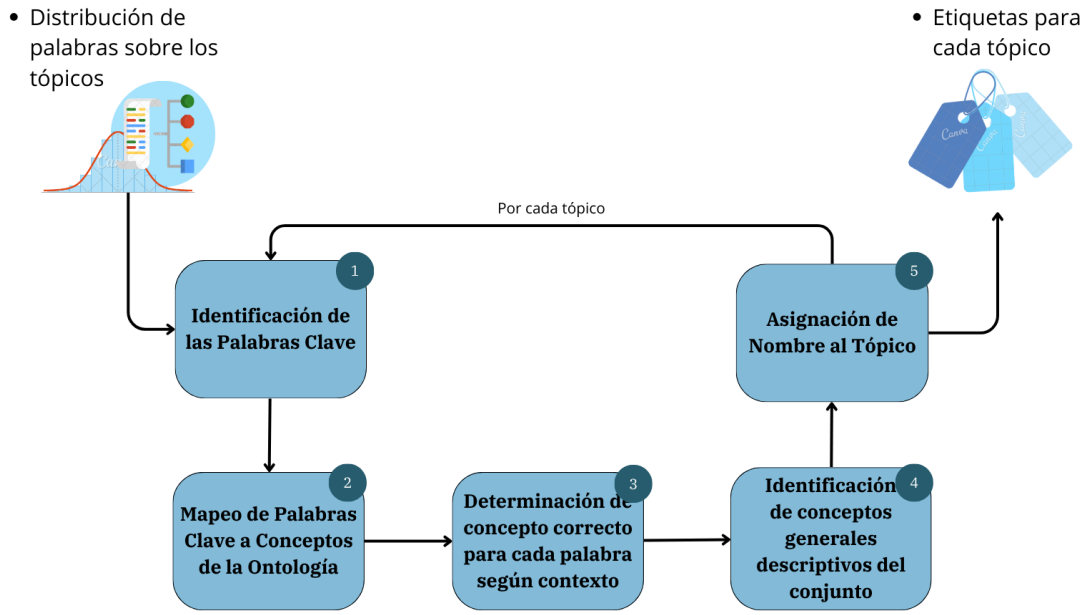

Figura 2.4: Asignación de nombres a tópicos

En la primera etapa, se seleccionan las palabras más probables para cada tópico a partir de las probabilidades proporcionadas por el modelo LDA. Este proceso incluye la aplicación de un umbral específico, lo que permite identificar de manera efectiva las palabras clave más representativas de cada temática. Estas palabras clave serán la base para la posterior asignación de nombres a los tópicos. Luego, cada palabra clave seleccionada se asigna a los conceptos pertinentes presentes en la ontología de

dominio general. Es importante tener en cuenta que las ontologías están diseñadas para abordar la polisemia del lenguaje natural, lo que significa que, en la mayoría de los casos, una palabra puede estar asociada a más de un concepto en la ontología.

Luego, se hace necesario en la siguiente etapa, seleccionar el concepto adecuado para cada palabra de acuerdo con el contexto en el que aparece en el corpus. Con este propósito se aplican técnicas de Disambiguación del Sentido de las Palabras (WSD, por sus siglas en inglés). Este paso es fundamental para garantizar una asignación precisa y contextualizada de conceptos ontológicos que engloben a las palabras clave identificadas en los tópicos. Se exploran tres variantes para abordar este desafío.

La primera variante representa una mejora significativa sobre el algoritmo tradicional de Lesk utilizado en WSD. Mientras que el algoritmo de Lesk convencional se basa en coincidencias exactas entre las palabras del contexto y las presentes en las definiciones de los sentidos, la variante propuesta supera esta limitación. El enfoque propuesto aprovecha embeddings de palabras para capturar la semántica y el significado contextual de las palabras, calculando la media de los embeddings del contexto. Además, para cada definición en la ontología, se calcula la media aritmética de los embeddings de las palabras asociadas a esa definición. La elección de la definición adecuada se realiza comparando las medias de embeddings. Se selecciona la definición cuya media de embeddings se aproxima más al contexto en términos de similitud del coseno. De esta forma se permite una mayor flexibilidad y capacidad para capturar relaciones semánticas más sutiles, mejorando así la precisión y el rendimiento del algoritmo en tareas de WSD.

Este problema de WSD se puede abordar también como un problema de optimización. Se plantea como la tarea de seleccionar, de una lista de n elementos (palabras), cada uno asociado a k características (definiciones), con k variable, n características exactamente, de modo que se maximice la similitud entre cada par de características seleccionadas. En este contexto, maximizar la similitud equivale a minimizar el camino entre dos definiciones en la ontología. Para resolver esta tarea, se utiliza el algoritmo Simplex y se concibe un algoritmo genético, permitiendo así encontrar la mejor combinación de conceptos ontológicos asociados a las palabras clave en los tópicos identificados. Este enfoque optimizado pretende mejorar la precisión en la asignación de conceptos y contribuir a una representación semántica más refinada de los tópicos en el contexto ontológico.

Una vez se haya seleccionado la definición correcta en la ontología para cada palabra clave, se procede a identificar los conceptos más generales asociados a cada una. Se busca construir una lista de conceptos generales para cada palabra y, posteriormente, se aplica una función de peso híbrida para determinar cuáles de estos conceptos describen mejor al grupo de palabras. Esta función de peso considera diversos factores para evaluar la relevancia de cada concepto general. Entre los factores se encuentran: la pofundidad en la jerarquía de la ontología, indicando cuán específico

es el concepto, la información de contenido para obtener contexto, la similitud con las palabras presentes en el contexto y la generalidad del concepto. La función de peso híbrida permite asignar una puntuación a cada concepto general, tomando en cuenta la combinación de estos factores. Finalmente, se toman los conceptos generales que han obtenido la puntuación más alta, lo que asegura una selección de los términos más adecuados y representativos para denominar los tópicos identificados. Este enfoque garantiza una asignación de nombres que considere tanto la estructura jerárquica de la ontología como la riqueza semántica del contexto circundante.

2.2. Recuperación por tópicos

En esta sección se aborda el diseño de un motor de búsqueda basado en modelos el modelo propuesto anteriormente. El propósito principal de este desarrollo es aprovechar los resultados obtenidos en la etapa anterior para mejorar y la eficiencia y visualización de los resultados en la exploración.

Esta etapa se inicia con la recepción de consultas por parte de los usuarios. Utilizando el modelo de tópicos LDA previamente entrenado, se identifican los tópicos más relevantes para la consulta. Posteriormente, se lleva a cabo un filtrado de documentos asociados a tópicos no relevantes para la consulta, mejorando significativamente la eficiencia del sistema al centrar la búsqueda únicamente en la información contextualmente relacionada con los temas de interés del usuario. Se calcula la similitud entre los documentos de los tópicos relevantes y la consulta, y se presentan los resultados al usuario ordenados segun su relevancia, por tópicos.

Capítulo 3

Implementación y Experimentación

En este capítulo, llevamos a cabo la materialización de la propuesta delineada en el capítulo anterior. Detallamos la implementación concreta de la metodología concebida, seguida de un análisis riguroso a través de experimentos diseñados para evaluar su desempeño y eficacia.

3.1. Implementación de la Metodología Propuesta

Describimos en detalle la traducción de los conceptos y estrategias de diseño en código ejecutable. Se proporcionan explicaciones paso a paso sobre cómo se llevó a cabo la implementación de cada componente, desde el preprocesamiento semántico hasta el proceso de recuperación por tópicos. Incluimos consideraciones técnicas relevantes, elecciones de herramientas y bibliotecas, y cualquier ajuste específico necesario para adaptar la metodología a la aplicación concreta.

3.2. Experimentación

Presentamos la estructura y diseño de los experimentos realizados para evaluar la eficacia de la metodología. Detallamos la selección de conjuntos de datos de prueba, establecemos métricas de evaluación pertinentes y explicamos las decisiones detrás de la configuración experimental. Este apartado proporciona un marco claro para la interpretación de los resultados y la validación de la propuesta.

3.3. Análisis de Resultados

Exponemos los resultados obtenidos a través de los experimentos y realizamos un análisis detallado de los mismos. Se comparan los rendimientos obtenidos con

otros enfoques existentes (si aplicable) y se evalúa la robustez de la metodología en diferentes condiciones. Además, destacamos cualquier hallazgo inesperado o patrones significativos identificados durante la experimentación.

Conclusiones

Conclusiones

Recomendaciones

Recomendaciones