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

25 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

1.1. Modelos de Tópicos

1.2. Ontologías

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.

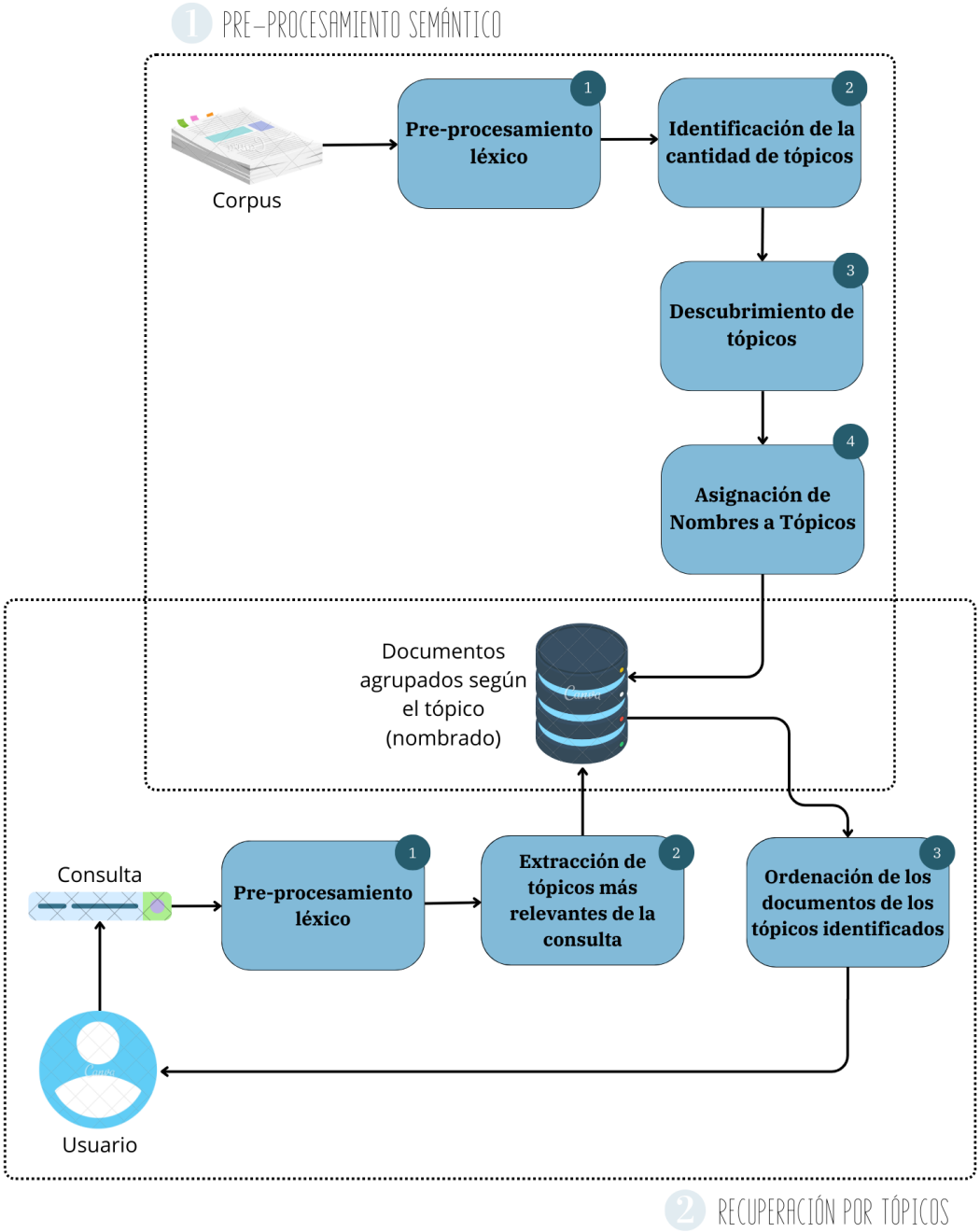

Figura 2.1: Arquitectura del sistema

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.

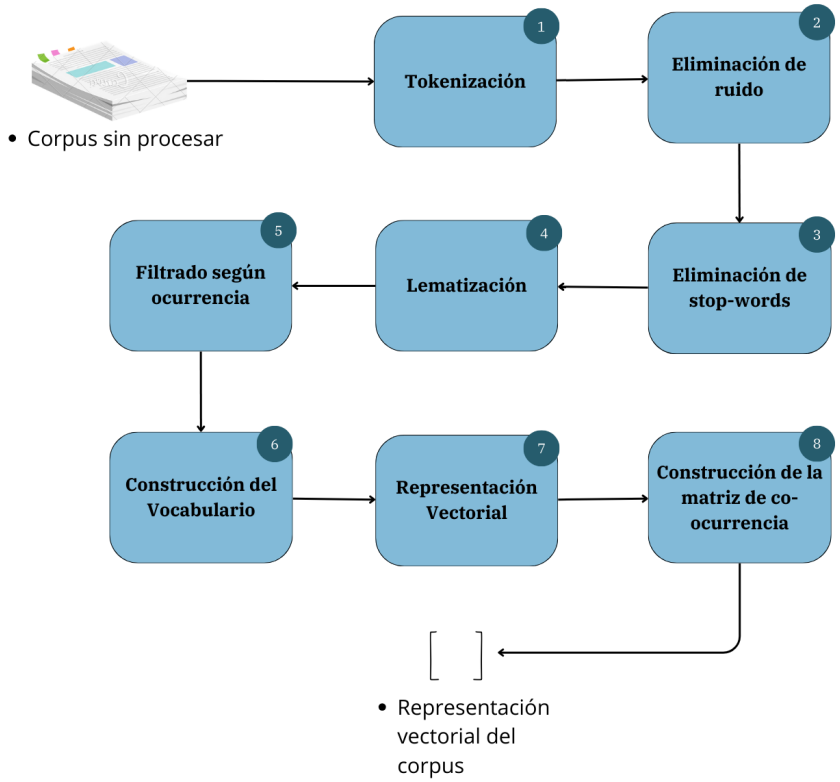

Figura 2.2: Pre-procesamiento léxico

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.

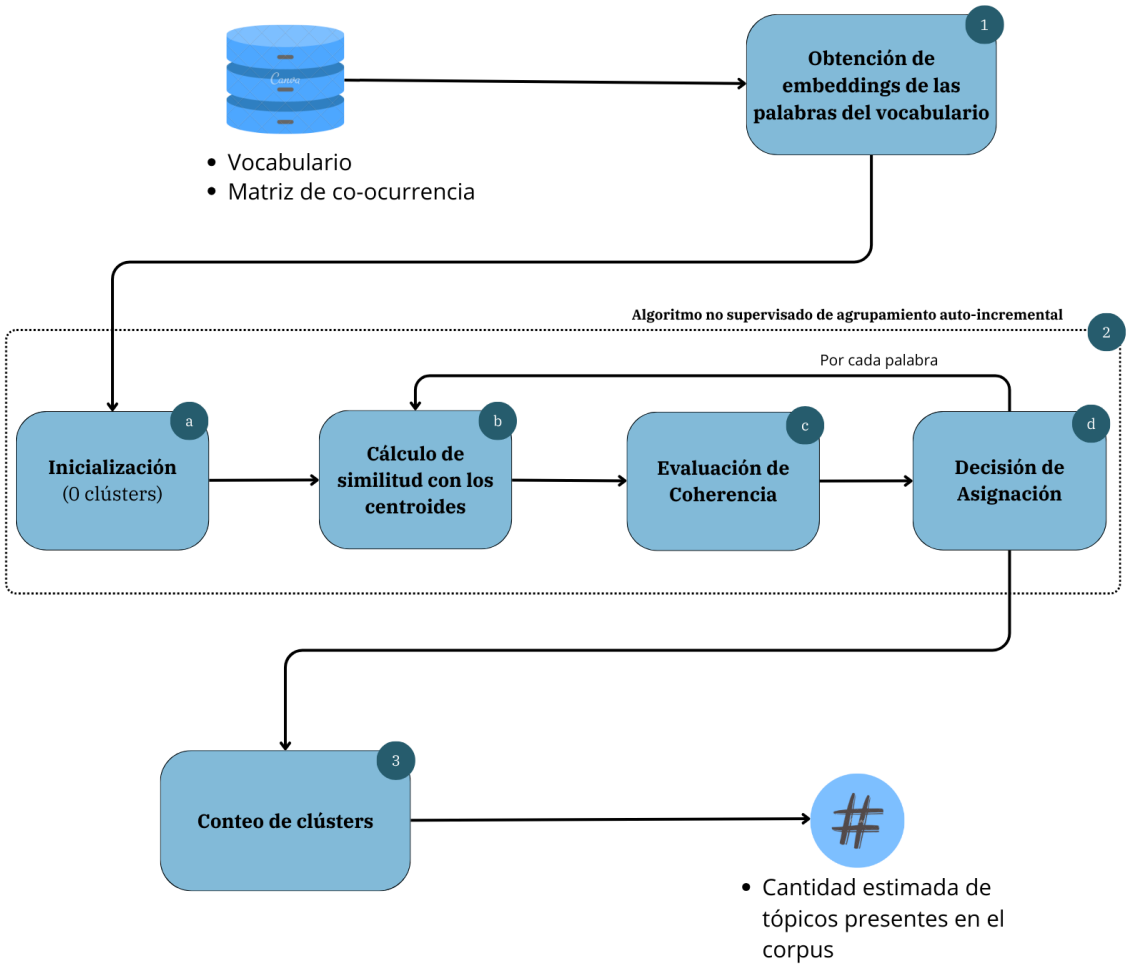

Figura 2.3: Identificación de la cantidad de tópicos

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).

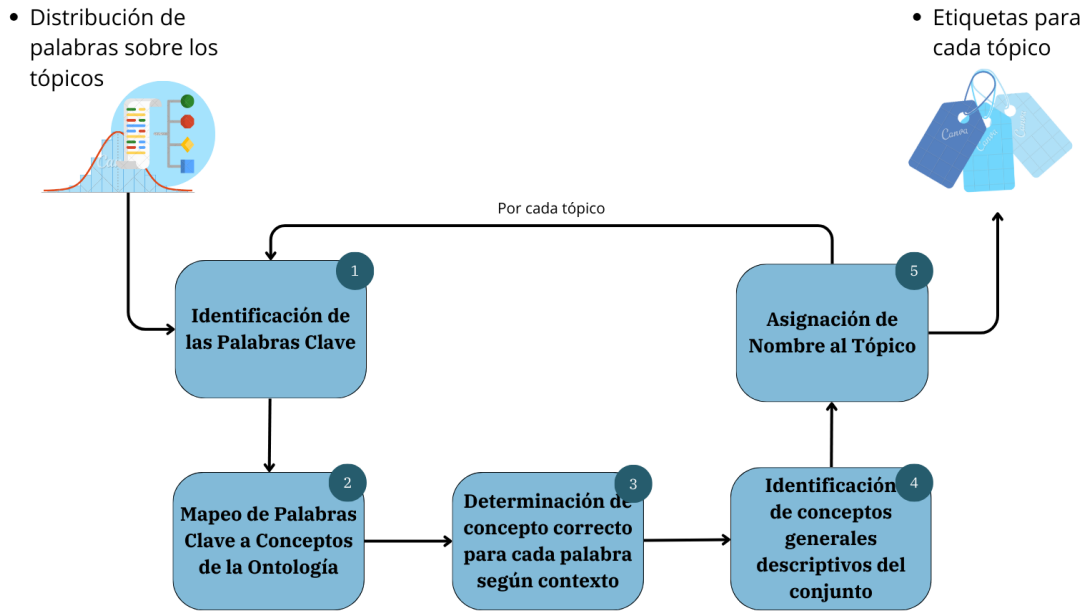

Figura 2.4: Asignación de nombres a tópicos

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