

Minería de textos orientada a la indización de artículos académicos

Text mining focused on indexing academic articles

Mariela Lucina Castro-Cáceres

Investigadora independiente

Contacto: mlcastrocaceres@gmail.com

<https://orcid.org/0000-0003-0949-0352>

Rubén Urbizagástegui-Alvarado

Universidad de California en Riverside, California

Contacto: ruben@ucr.edu

<https://orcid.org/0000-0001-5014-801X>

Resumen

Analiza el potencial de la minería de textos para apoyar la indización documental con el fin de facilitar una recuperación eficiente de parte del usuario en un sistema de información. Como unidad de análisis se utilizó el texto completo de un artículo científico, el cual fue guardado como texto plano en Notepad++. A continuación, se procedió a la preparación de los datos, los cuales fueron procesados en R/RStudio con diversos paquetes para la minería de textos, el análisis de frecuencias y la visualización de asociaciones de palabras. La comparación entre términos asignados por el autor y los generados automáticamente se llevó a cabo mediante el índice de similitud

y disimilitud de Jaccard. Los resultados muestran que la minería de textos identifica un conjunto más amplio de términos potencialmente útiles para la indización, con solo un 12,5% de coincidencia entre las palabras del documento y las sugeridas por el software, lo que evidencia una baja similitud entre ambos. Se concluye que la minería de textos constituye un complemento relevante de la indización tradicional y aporta insumos para la revisión y actualización de tesauros y vocabularios controlados.

Palabras clave: Minería de textos, indización, palabras clave, Índice de Jaccard, análisis de contenido

Abstract

This study analyzes the potential of text mining to support document indexing in order to facilitate efficient user retrieval within an information system. The full text of a scientific article, saved as plain text in Notepad++, was used as the unit of analysis. The data was then prepared and processed in R/RStudio using various packages for text mining, frequency analysis, and word association visualization. The comparison between terms assigned by the author and those automatically generated was performed using the Jaccard similarity-dissimilarity index. The results show that text mining identifies a broader set of potentially useful terms for indexing, with only a 12.5% match between the words in the document and those suggested

by the software, indicating a low level of similarity between the two. The study concludes that text mining is a relevant complement to traditional indexing and provides input for the revision and updating of thesauri and controlled vocabularies.

Keywords: Text mining, indexing, keywords, Jaccard Index, content analysis

Recibido: 16/03/2026 / Aceptado: 26/05/2026

Introducción

Los conocimientos se presentan estructurados en diversos formatos: impresos (revistas, libros, capítulos de libros, folletos, avances de investigación, informes técnicos, normas técnicas, patentes, etc.); imágenes (fotografías, planos, mapas, dibujos, etc.); sonoros (películas, discos compactos, discos de vinilo, etc.); y multimedia (como una combinación de imágenes, sonidos y otros textos). Históricamente, la forma de representación del conocimiento ha ido cambiado de soporte de almacenamiento y ahora

encontramos el conocimiento registrado tanto en soportes tradicionales como el papel, como en soportes electrónicos, ópticos y magnéticos. Este aumento del conocimiento se traduce en la disponibilidad de una enorme cantidad de información, recurso que está siendo considerado el capital más importante de nuestra sociedad, ya que representa un alto valor añadido para el sector productivo industrial y empresarial” (Pinto, 2000, p. 64, traducción propia).

La enorme cantidad de información reunida en un centro de información y/o documentación destinada a satisfacer las demandas

de los usuarios necesita pasar por un proceso de organización de la información y del conocimiento. Uno de los pasos fundamentales de este proceso es la indización, orientada a identificar y describir el contenido de un documento a través de términos lingüísticos que lo representen de manera adecuada. En este proceso, la lectura del documento es el primer procedimiento de análisis, por lo tanto, la lectura se convierte en un paso decisivo, un proceso complejo que implica una formación intelectual apropiada, capacidad de concentración y habilidades para la síntesis de parte del indizador. Por tanto, la indización es un conjunto de actividades que consiste en identificar en los documentos sus

huellas descriptivas o macro-proposiciones y luego extraer los elementos descriptores (sintagmas) indicadores de su contenido para su posterior recuperación. Estos descriptores constituirán la representación de los elementos que indican el contenido del documento y no su representación, ya que ésta sólo puede ser representada por el propio documento” (Jean-Claude Gardin, 1974 citado por Pinto, 2000, p. 65, traducción propia).

El procesamiento manual del conocimiento contenido en los textos es costoso, laborioso y exige una extensa lectura y análisis de los documentos por parte de los organizadores de la información. Sin embargo, en este proceso de indización, los especialistas pueden hacer uso de una combinación humana y mecánica conocida como indización semiautomática o asistida por computadora. En una primera fase, el algoritmo efectúa automáticamente la indización de los documentos, considerando las ocurrencias de las palabras más frecuentes en el texto. En una segunda fase, el indexador humano refina la lista de descriptores generada por el algoritmo, realizando los ajustes pertinentes y recurriendo al conjunto de "autoridades" construidas para asistir en ese proceso, como lo son los diversos tipos y clases de registros de autoridad y los vocabularios controlados.

La exploración de este proceso de indización semiautomática constituye el objeto de estudio de este artículo, el cual se orienta a examinar el proceso de extracción y construcción de palabras clave a partir de un texto previamente seleccionado, mediante la aplicación de un método no supervisado. No obstante, es necesario señalar que las actividades de organización y representación del conocimiento deben adaptarse y responder al impacto de las nuevas tecnologías disponibles, haciendo posible un diálogo creativo y positivo entre la generación de conocimientos

y el uso de la información destinada a la reproducción de nuevos conocimientos. En esta línea, el desarrollo del artículo se sustenta en la perspectiva de la minería de textos aplicada a la recuperación de términos de información; en particular, se propone utilizar el texto académico como un objeto desde el cual es posible recuperar la terminología empleada por el autor con el fin de construir vocabularios controlados que permitan una recuperación más precisa y significativa de la información.

En un documento académico publicado, generalmente la palabra clave es atribuida por el autor, pero utilizando unidades léxicas libres, buscando así representar sintéticamente el contenido temático del texto. El autor tiene el dominio informacional sobre el texto, pues es su creador. Se entiende entonces que la palabra clave debe ser un término del área de conocimiento del autor y, por lo tanto, una unidad de representación como de recuperación de la información contenida en el documento. Se propone establecer una representación del contenido, con el fin de facilitar una recuperación eficiente de parte del usuario en un sistema de información. Para ello, es necesario que exista una convergencia entre el lenguaje del autor expresado en su documento publicado, el lenguaje del usuario que busca la información y el lenguaje utilizado por el organizador de la información para el tratamiento

temático del documento en un sistema de información.

Con el fin de alcanzar el objetivo planteado, este documento se organiza de la siguiente manera: después de una breve introducción y establecimiento del objetivo del artículo, se presenta un marco teórico sobre la minería de textos aplicada a los estudios relacionados con la ciencia de la información y la documentación. A continuación, se describe la metodología, esto es, las unidades de análisis, la forma de recolección de los datos y el modo de medición de estos; seguidamente, se presentan los resultados obtenidos y se exponen las conclusiones. Finalmente, se incluye la bibliografía utilizada para la elaboración de esta investigación.

Revisión de la literatura

Identificando una lista de las palabras más frecuentemente utilizadas en un documento, la minería de textos puede ser empleada como instrumento para la construcción de vocabularios controlados que pueden apoyar la construcción de ontologías (Araújo Júnior y Tarapanoff, 2006). Buscando esas posibilidades Rezende; Marcacini & Moura (2011) describen un proceso de extracción y organización del conocimiento a través de métodos no supervisados. En primer lugar, exponen las principales técnicas de preprocesamiento de documentos con el objetivo de obtener una representación concisa y estructurada

del acervo textual. A continuación, discuten los principales algoritmos de agrupación de textos, así como las técnicas para seleccionar descriptores para los grupos formados. Seguidamente, presentan índices de validación de conglomerados, discutiendo su importancia para la evaluación del conocimiento extraído. Finalmente, ilustran algunos ejemplos y aplicaciones que se benefician de un proceso de minería de textos con el fin de mostrar la utilidad práctica de la metodología descrita a lo largo del artículo.

Asimismo, Bezerra y Guimaraes (2014) aplicaron la minería de textos para identificar los términos más utilizados en los resúmenes de investigaciones publicados entre 2003 y 2012 en las bases de datos nacionales e internacionales: EBSCO, Web of Knowledge, Scopus, ScienceDirect y SciELO Brasil. En los resultados, los autores expusieron que los temas relacionados con prácticas, procesos, sistemas gestión del conocimiento y tecnologías de la información aparecen de manera recurrente en los resúmenes, tanto nacionales como internacionales. Por otro lado, los artículos nacionales se distinguen por presentar términos más asociados a aspectos abstractos de la gestión del conocimiento (modelos, conocimiento tácito y organizacional), mientras que los artículos internacionales se caracterizan por términos más pragmáticos (eficacia, implementación, relaciones y desarrollo). En ese sentido, la

minería de textos puede ayudar a determinar posibles tendencias de investigación en el área de gestión del conocimiento, contribuyendo así al establecimiento de estudios más alineados con el contexto más actual.

Silva y Souza (2014) afirman que el texto escrito es una forma importante de registrar información y que gran parte de este contenido informativo se encuentra disponible en medios digitales. No obstante, las computadoras tratan el texto como una cadena de caracteres que no tienen significado. El área de Procesamiento del Lenguaje Natural (PNL) se esfuerza por extraer significados del texto. Por ello, los autores presentan una revisión de este tema y proponen un método automatizado que utiliza una heurística determinista llamada Heudet, que tiene como objetivo extraer bigramas de los textos. El principal propósito perseguido es extraer el significado del texto a través de un conjunto de expresiones de varias palabras identificadas. Los autores demostraron empíricamente que el procesamiento determinista presentó ganancias de precisión en relación con las otras técnicas estadísticas proporcionadas por el paquete NSP. En consecuencia, Heudet se puede utilizar con ventaja como parte de un procesamiento de lenguaje natural que requiere la identificación de expresiones de varias palabras en un documento.

Con el fin de identificar tendencias de corto, mediano y largo plazo de

las ciencias de la información en Brasil, Trucolo & Digiampietri (2014) realizaron un metaestudio para desarrollar y aplicar una metodología para identificar tendencias en temas y líneas de investigación, a partir de la información ofrecida por los currículos de estudio registrados en la plataforma Lattes de los doctores que trabajan en dicho campo. A partir de la extracción automática de términos, seleccionaron los tres términos más importantes para ejemplificar el comportamiento temporal de cada uno de ellos. Los autores encontraron que los términos "ciencia de la información" y "gestión del conocimiento" tuvieron un crecimiento similar hasta 2006, cuando se separaron. Mientras que el término "ciencia de la información" siguió creciendo hasta 2009 cuando se estabilizó y tuvo una caída importante en 2012. La palabra clave "gestión del conocimiento" comenzó a declinar en 2006 y se estabilizó a partir de 2008. El término "medio ambiente", a diferencia de los otros dos términos, tuvo un comportamiento estable con alzas y bajas no muy significativas durante el período estudiado.

Araújo *et al.* (2016) aplicaron técnicas de minería de textos a la literatura sobre esquistosomiasis publicada entre 2005 y 2015, disponible en la colección *Memórias do Instituto Oswaldo Cruz*. Estudiaron 179 resúmenes de artículos científicos seleccionados mediante el descriptor "schisto". A continuación, analizaron esa literatura con el software Weka

3.6.12; posteriormente, extrajeron los términos más frecuentes mediante el método *Term Frequency-Inverse Document Frequency* (TF-IDF). Además, categorizaron los documentos utilizando técnicas de aprendizaje automático basadas en árboles de decisión y clasificación de clústeres. Finalmente, señalaron que el conocimiento descubierto en el estudio ofrece una nueva dimensión para abordar y comprender el contenido del acervo textual sobre esquistosomiasis.

Una investigación exploratoria y experimental que evalúa la calidad de la indexación automática a través de un índice de consistencia en el campo de la ciencia de la información en Brasil fue realizada por Bandim y Correa (2019). El experimento consistió en la indización automática del corpus de 60 artículos científicos seleccionados, mediante la aplicación del Sistema de Indización Semi-Automática (SISA) con los descriptores del Tesauro Brasileño de Ciencia de la Información. A continuación, compararon los términos asignados por SISA con las palabras clave de los autores, utilizando el criterio de consistencia relajada, y obtuvieron un índice medio de consistencia del 14% con la asignación de términos generales habilitada en la configuración SISA, y del 19% sin habilitar dicha asignación. En consecuencia, los autores concluyeron que la indización automática tuvo un desempeño satisfactorio, medido a partir del índice de consistencia.

Silva y Correa (2020) comparan dos sistemas de indización automática por asignación multilingüe: SISA y MAUI. El Sistema de Indización Semi-Automática (SISA) desarrollado en España, propuesto inicialmente para el área de biblioteconomía y documentación, es un sistema que realiza la indización automáticamente siguiendo un tesoro y reglas de indización predeterminadas, basadas en la frecuencia y posición de los términos. Por su parte, *Multi-purpose Automatic Topic Indexing* (MAUI), sistema de origen neozelandés, utiliza un tesoro y un algoritmo de aprendizaje automático para generar un modelo a partir de los resultados de indización intelectual, con términos representados mediante características estadísticas. En las conclusiones, los autores afirman que MAUI es más prometedor por los recursos adicionales de procesamiento de lenguaje natural que implementa, y por presentar la especificidad de entrenar un modelo de indización por medio de un algoritmo de aprendizaje automático. El uso de técnicas de aprendizaje automático parece ofrecer una mayor eficiencia en la representación de la información, al generar un modelo a partir de la indización manual, considerada por los investigadores del área como la forma de referencia o estándar de calidad en la indización.

Finalmente, Ferreira y Correa (2021) llevaron a cabo una investigación de naturaleza cualicuantitativa, de

carácter exploratorio, con el propósito de ofrecer una visión general de los artículos sobre minería de textos publicados en revistas brasileñas de ciencias de la información. El estudio se basó en una investigación bibliográfica sustentada en los procedimientos de estudios métricos y análisis de contenido. Los autores, realizaron una búsqueda con la palabra clave "*Text Mining*" en las bases de datos Brapci y SciELO, obteniendo 23 documentos del campo de las Ciencias de la Información que contenían el término en alguna parte de los metadatos.

En una segunda búsqueda, usando el término compuesto "Descubrimiento de conocimiento" recuperaron 14 registros. Posteriormente, combinaron los dos conjuntos de documentos y, tras eliminar los artículos duplicados, obtuvieron un total de 28 documentos. Después de los filtrados correspondientes, analizaron 14 registros del área de Ciencias de la Información que contenían el término en uno de los campos de los metadatos. A lo largo de 18 años, sólo 28 trabajos fueron identificados como investigaciones que declaraban explícitamente el propósito de estudiar el descubrimiento del conocimiento a través de la minería de textos. No obstante, 13 de estos documentos abordaban específicamente la minería de textos científicos. Otro aspecto relevante identificado por los autores es que la mayoría de los artículos fueron publicados a partir de 2014,

lo que indica que se trata de un tema reciente y de creciente interés de investigación en las ciencias de la información en Brasil.

Como se desprende de la literatura revisada, no parece existir un acercamiento al proyecto R ni una exploración de la minería de textos que aproveche la multiplicidad de paquetes que este proyecto ofrece para dicho fin. En ese sentido, podría resultar oportuno experimentar con algunos de estos paquetes para crear terminologías de indización semiautomática, con el propósito de obtener una terminología más ajustada al vocabulario utilizado por los propios autores.

Material y método

Como unidad de análisis experimental se tomó el texto completo del artículo de Jaume *et al.* (2021), quienes discuten la teoría epistémica laica y el enfoque cognitivo-motivacional del conocimiento. Se copió el texto completo, el cual se guardó en formato de texto plano con Notepad++ y codificación UTF-8. Inicialmente, se descartaron el resumen y *abstract* en inglés, así como las referencias bibliográficas incluidas al final del texto. A continuación, se procedió a la preparación de los datos, con la finalidad de evitar que los algoritmos de la minería de textos separarán los nombres propios mencionados en el texto. Posteriormente, se realizó la limpieza de la información, eliminando signos de puntuación,

fechas, números, espacios en blanco innecesarios, y se convirtió todo el texto a minúsculas.

Asimismo, se eliminaron artículos, adverbios, entre otras palabras que no aportan significado en este contexto específico, en particular los denominados *stopwords*. Todo este proceso se realizó utilizando los paquetes *tm* (Feinerer y Hornik, 2020) y NLP (Hornik, 2020). Para la remoción de los *stopwords* se utilizó el paquete *SnowballC* (Bouchet-Valat, 2020). Para la visualización de las frecuencias de los términos identificados se utilizaron los paquetes *ggplot2* (Wickham, 2016) y *ggthemes* (Arnold, 2021). En el trazado de las asociaciones entre las palabras se utilizó el paquete *textplot* (Wijffels, 2021). Todos estos paquetes

se emplearon en la versión 4.0.5 de R y 1.4.1 de *RStudio*, *softwares* seleccionados porque son gratuitos, que contienen recursos para el análisis de minería de textos y ofrecen numerosas funciones que pueden ampliarse mediante la incorporación de nuevos paquetes.

Medición de los datos

Para la medición de los datos se utilizó el índice de similitud de Jaccard, el cual establece una medida de la similitud entre dos conjuntos de datos. Este índice varía de 0 a 1: cuanto más cercano a 1, más similares son los dos conjuntos de datos y, cuanto más próximo a 0, más disímiles son los datos. El índice de similitud de Jaccard se calcula de la siguiente manera:

$$\text{Índice de Jaccard} = \frac{\text{número de observaciones comunes en ambos conjuntos}}{\text{total de observaciones en ambos conjuntos}}$$

Escrito en forma de notación es:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

Si dos conjuntos de datos comparten exactamente los mismos miembros, el índice de similitud de Jaccard será 1. Caso contrario, si no tienen datos en común, su similitud será 0. Por ejemplo, si las palabras clave del texto analizado son correctas, se espera que el paquete *tm* construya palabras clave que

coincidan plenamente con las palabras clave ofrecidas por el autor del texto; por lo tanto, el índice de Jaccard será igual a 1 (plena coincidencia). Si este índice es bajo (cercano a cero) significará que existe disimilitud y que el paquete *tm* está aportando otras palabras clave que bien podrían ser consideradas para mejorar la

recuperación del contenido del texto. Por otro lado, se tiene el índice de

disimilitud o distancia de Jaccard. Este índice es representado como:

$$Disimilitud = 1 - \text{Índice de similitud}$$

Este índice también varía de 0 a 1. Cuanto más cercano a 0 (cero) son los valores más similares son las palabras clave y cuando más cercanas a 1 más disímiles serán las palabras

clave construidas. Ambos índices de Jaccard fueron estimados utilizando *RStudio*, con el Proyecto R en su versión 4.1.3.

Resultados

Mediante la aplicación del paquete *tm* se obtuvieron las palabras clave listadas en la Tabla 1. Con este resultado se pueden construir las siguiente palabras clave: necesidad

de cierre, cierre específico, teoría epistémica laica, actividad epistémica, bloques de lego, cierre inespecífico, y motivación epistémica.

Tabla 1

Palabras clave construidas por el paquete tm

PC del documento original	PC construidas con el paquete "tm"	Frecuencia
Epistemología	necesidad cierre	12
Teoría epistémica laica	cierre específico	8
Necesidad de cierre cognitivo	evitar cierre	8
	teoría epistémica	8
	epistémica laica	7
	actividad epistémica	6
	bloques lego	6
	cierre inespecífico	5
	motivación epistémica	5

Nota. Fuente: Elaboración propia, 2025

El Índice de Jaccard estimado fue de 0.125 indicando una baja similitud entre las palabras clave construidas por el autor del documento y las construidas por el paquete *tm*. En otras palabras, solo hay coincidencia en 12.5% del total de palabras. Por lo tanto, el índice de disimilitud es igual a 0.875. Eso significa que el paquete *tm* contribuye con 0.875 (87.5%) de otras palabras clave que bien podrían ser consideradas para garantizar la recuperación de este documento. La tarea de decidir cuáles serán las palabras clave escogidas es responsabilidad del indizador, quien, en su toma de decisiones, debe basarse en su familiaridad con el área, en su experiencia en la construcción de autoridades de asuntos y en el lenguaje utilizado por sus usuarios locales.

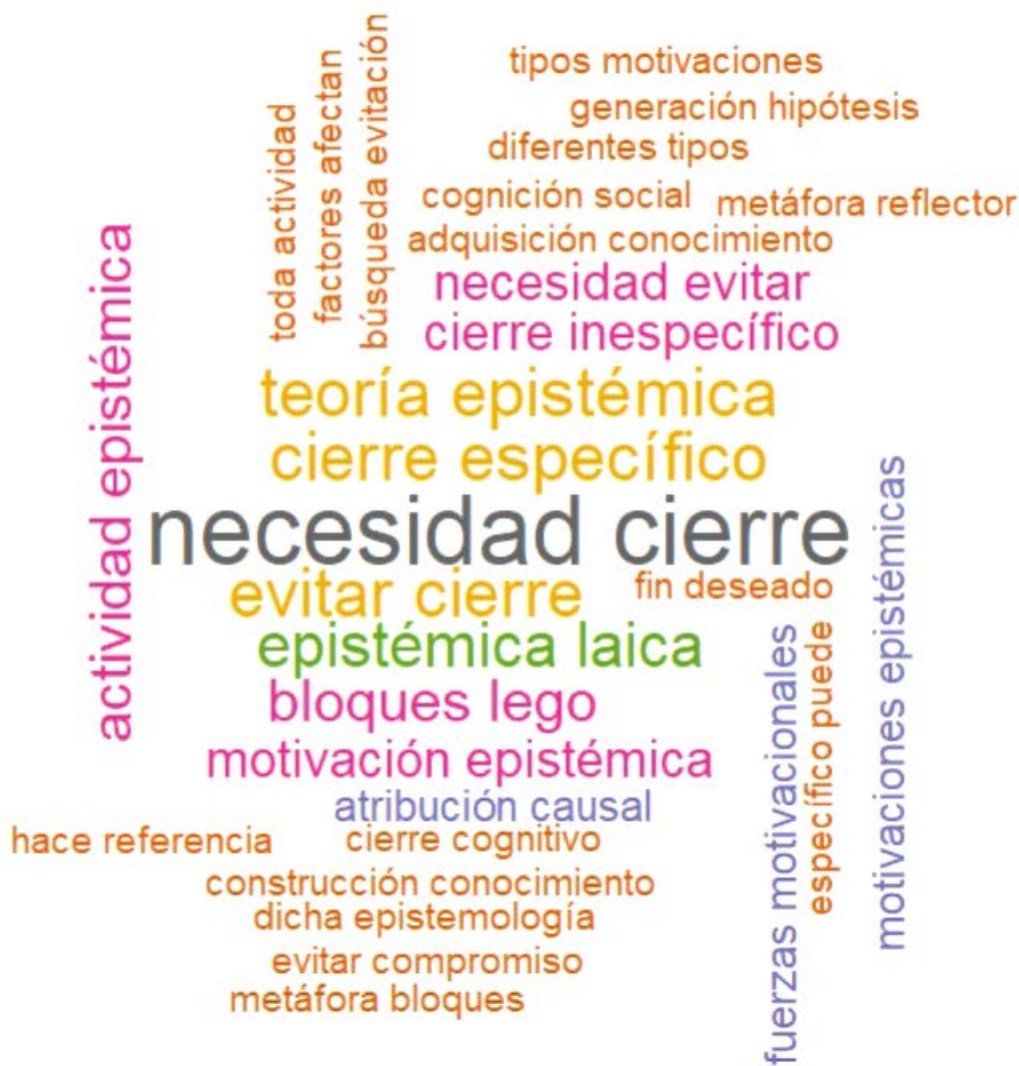
Los autores del documento bajo análisis propusieron las palabras clave: epistemología / teoría epistémica laica / necesidad de cierre cognitivo. La "teoría epistémica laica" se conceptualiza como un modelo teórico que aborda, por un lado, el proceso mediante el cual las personas adquieren conocimientos sobre sí mismas y sobre el mundo que las rodea y, por otro, describe los aspectos invariantes que rigen la formación de creencias de diversos ámbitos (Kruglanski, 1989). Propone que, en todo proceso de conocimiento subyace un deseo de alcanzar ciertos estados epistémicos a fin de abandonar estados de ambigüedad o incertidumbre.

Este deseo epistémico se denomina "necesidad de cierre cognitivo", término psicológico que describe el deseo del ser humano por obtener una respuesta concreta a cualquier cuestión específica, sin dejar espacio a la confusión o a la ambigüedad; por esta razón, se destaca la relevancia de los estudios de la necesidad de cierre cognitivo en la toma de decisiones personales (Recart et al., 2021; Moyano y Trujillo, 2018). En este contexto, la "actividad epistémica" se refiere a las acciones que realizan los sujetos al intentar entender por qué las personas actúan de la forma en que lo hacen y, a su vez, el porqué de sus propias conductas; de esta manera, toda actividad epistémica promueve de cierta manera la producción de conocimiento.

En todo caso, con el paquete *tm* se puede construir una nube de palabras que ofrece una variedad de términos que pueden servir como insumo para seleccionar las palabras clave más adecuadas para el texto (Figura 1). Se observa que "necesidad de cierre", "teoría epistémica", "cierre específico", "epistémica laica" y "motivación epistémica", aparecen como fuertes candidatas para ser consideradas como palabras clave.

Figura 1

Nube de palabras de bigramas



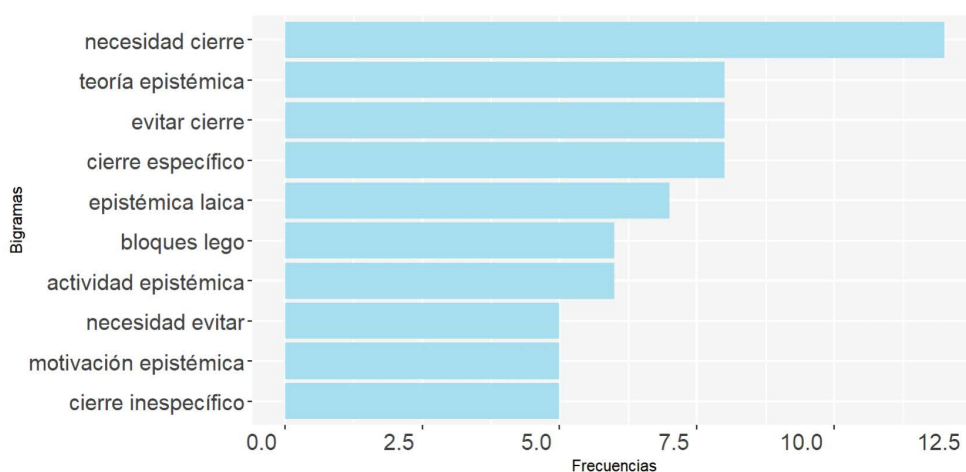
Nota. Fuente: Elaboración propia, 2025

Asimismo, dicho paquete además de mostrar las palabras clave en trigramas (Tabla 2), permite generar gráficos de frecuencia de bigramas identificados en el texto (Figura 2), los

cuales contribuyen a corroborar las opciones de selección de las palabras clave que mejor lo describen. Las expresiones se confirman como candidatas a palabras clave.

Figura 2

Bigramas utilizados en el texto



Nota. Fuente: Elaboración propia, 2025

Tabla 2

Palabras clave construidas por el paquete tm

PC del documento original	PC construidas con el paquete "tm"	Frecuencia
Epistemología	Teoría epistémica laica	7
Teoría epistémica laica	Necesidad evitar cierre	5
Necesidad de cierre cognitivo	Necesidad cierre específico	4
	Cierre específico puede	3
	Evitar cierre específico	3
	Metáfora bloques lego	3
	Necesidad cierre inespecífico	3

Nota. Fuente: Elaboración propia, 2025

Asimismo, con el paquete *tm*, se pueden elaborar los trigramas, los cuales se pueden representar, tanto en forma de gráfico de frecuencia (Figura 4), como mediante una nube

de palabras (Figura 3), lo que ofrece información adicional para afinar la elección de las palabras clave más pertinentes para el texto.

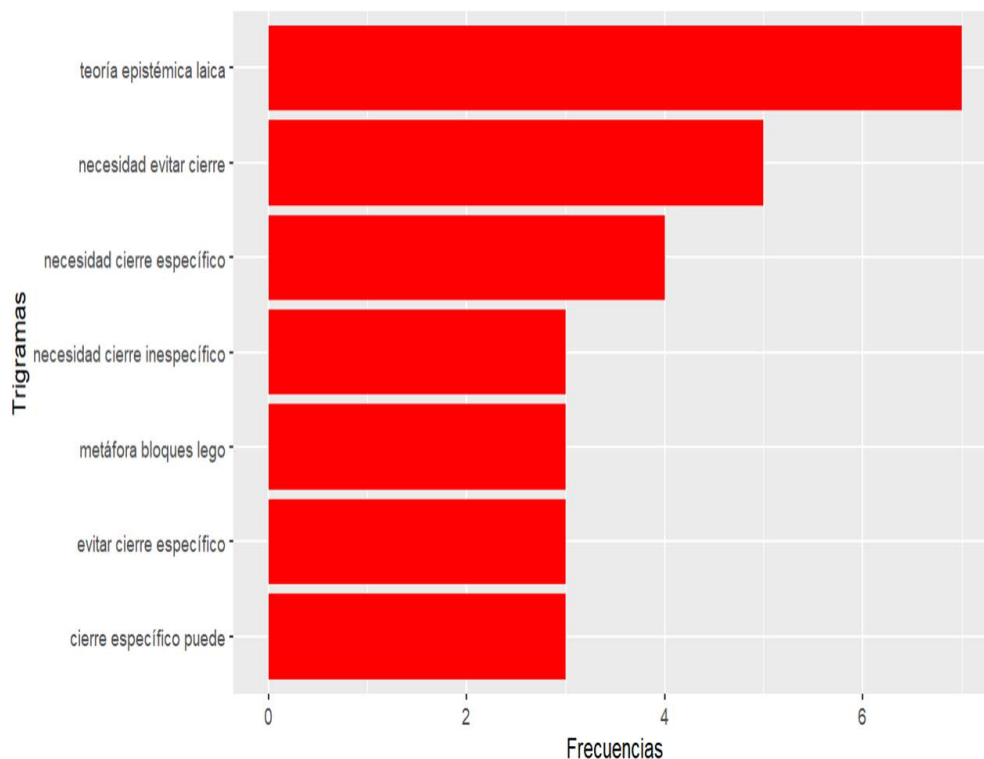
Figura 3

Trigramas utilizados en el texto



Nota. Fuente: Elaboración propia, 2025

Figura 4

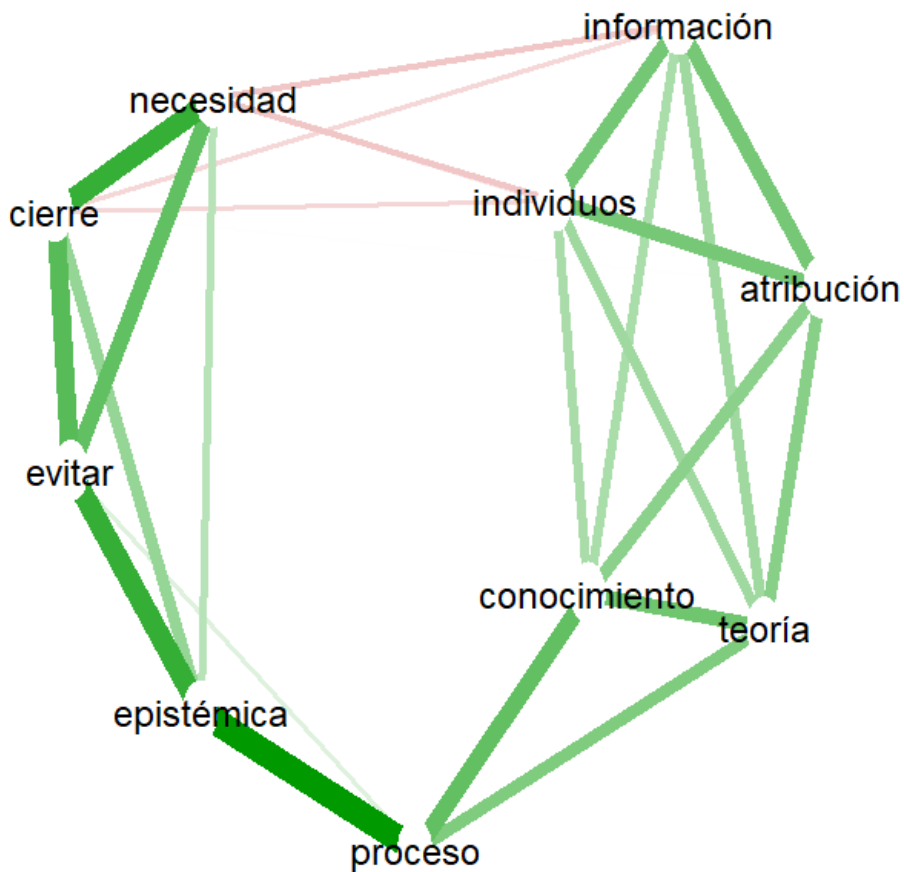
Frecuencia de trigramas de palabras

Nota. Fuente: Elaboración propia, 2025

Finalmente, con una combinación del paquete *tm* y el paquete *textplot* se puede producir una red de palabras (Figura 5), las cuales pueden complementar muy bien las posibilidades de selección de

las palabras clave. Por ejemplo, se puede jugar con: necesidad de evitar el cierre, teoría del proceso de conocimiento, proceso epistémico, atribución de individuos, información en individuos, etc.

Figura 5

Red de palabras clave*Nota.* Fuente: Elaboración propia, 2025

Conclusiones

Las palabras clave otorgadas a las publicaciones académicas son esenciales para el proceso de difusión y comunicación de la actividad científica, ya que representan, de forma condensada, el contenido temático de esas publicaciones. Representar adecuadamente los asuntos contenidos en esos documentos es fundamental para la recuperación de la información; los términos de indización o palabras clave construidos y asociados a esos documentos son los productos que garantizarán su recuperación en las búsquedas realizadas por los usuarios interesados en estos asuntos.

Estas palabras clave también pueden servir como insumo para la elaboración de tesauros y vocabularios controlados, utilizando la minería de textos como soporte para la actualización de estos instrumentos. Finalmente, la incorporación de la minería de textos en el campo de la bibliotecología y la ciencia de la información obliga a repensar el rol del profesional dedicado a la organización de la información y del conocimiento. Su participación en la construcción de la representación documental no debería limitarse a reproducir encabezamientos de materia, descriptores y palabras clave elaborados para otras realidades históricas y contextos sociales, ni ser el difusor de influencias funcionalistas. Más bien, está llamado a enriquecer y

ampliar esos repertorios, asumiendo un papel activo en el diseño de vocabularios controlados pertinentes para su entorno histórico local, regional o nacional.

La minería de textos, como área de estudio orientada al procesamiento de documentos textuales, desarrolla y aplica métodos que facilitan la extracción de conocimientos útiles para la organización de la información. En este estudio se comprobó, mediante el índice de Jaccard, que los autores y editores no siempre seleccionan las palabras clave que mejor representan el contenido de un texto, lo que puede limitar su adecuada recuperación en los sistemas de información.

Si bien en la actualidad existe una amplia oferta de herramientas para la extracción de palabras clave y ésta se ha incrementado con la adopción masiva de la inteligencia artificial, muchas de estas soluciones presentan limitaciones en términos de precisión y exhaustividad y, en sus versiones gratuitas, incorporan además restricciones de uso (como el número de documentos procesables, la longitud del texto o el volumen de consultas), ofreciendo, en consecuencia, resultados menos fiables que los servicios de pago mejor mantenidos y evaluados. En ese contexto, el empleo de *software* libre, como el Proyecto R utilizado en

este estudio, se presenta como una buena alternativa, ya que combina gratuidad, resultados satisfactorios, transparencia en los procedimientos y posibilidades de adaptación a diversas necesidades, lo que constituye una ventaja significativa para aquellas instituciones que cuentan con recursos limitados.

Mariela Lucina Castro-Cáceres - Rubén Urbizagástegui-Alvarado

Referencias

- Araújo, D. A. de O., David, L. R. de S., Rios, R. S. H., & Veloso, R. R. (2016). Descoberta de conhecimentos sobre a esquistossomose a partir de documentos científicos utilizando técnicas de mineração de texto. *Pesquisa Brasileira em Ciência da Informação e Biblioteconomia*, 11(2), 173–186. <https://doi.org/10.22478/ufpb.1981-0695.2016v11n2.31846>
- Araújo Junior, R. H., & Tarapanoff, K. (2006). Precisão no processo de busca e recuperação da informação: Uso da mineração de textos. *Ciência da Informação*, 35(3), 236–247. <https://doi.org/10.1590/S0100-19652006000300023>
- Bandim, M. A. S., & Correa, R. F. (2019). Indexação automática por atribuição de artigos científicos em português da área de Ciência da Informação. *Transinformação*, 31, 1–12. <https://periodicos.puc-campinas.edu.br/transinfo/article/view/5921>
- Bezerra, C. A., & Guimarães, A. J. R. (2014). Mineração de texto aplicada às publicações científicas sobre gestão do conhecimento no período de 2003 a 2012. *Perspectivas em Ciência da Informação*, 19(2), 131–146. <https://doi.org/10.1590/1981-5344/2292>
- Ferreira, M. H. W., & Correa, R. F. (2021). Mineração de textos científicos: Análise de artigos de periódicos científicos brasileiros da área de Ciência da Informação. *Em Questão*, 27(1), 237–262. <https://doi.org/10.19132/1808-5245271.237-262>
- Jaume, L. C., Roca, M., Passero, M., & Azzollini, S. (2021). La teoría epistémica laica y su enfoque cognitivo-motivacional del conocimiento. *Persona*, 24(2), 11–26. [https://doi.org/10.26439/persona2021.n024\(2\).5438](https://doi.org/10.26439/persona2021.n024(2).5438)
- Kruglanski, A. W. (1989). *Lay epistemics and human knowledge*. Springer.
- Moyano, M., & Trujillo, H. (2018). Pérdida de significado, necesidad de cierre cognitivo y extremismo. *Ciencia Cognitiva*, 12(2), 45–47. <https://www.cienciacognitiva.org/files/2018-4.pdf>
- Pinto, V. B. (2000). Indexação documentária: Uma forma de representação do conhecimento registrado. *Revista de Letras*, 22(1–2), 62–69. https://repositorio.ufc.br/bitstream/riufc/16564/1/2000_art_vbpinto.pdf

- Recart, E., Jaume, L. C., Passero, M. T., & Azzollini, S. (2021). Efectos de la necesidad de cierre cognitivo en los estilos decisorios: Un abordaje exploratorio. *PSOCIAL*, 7(2), 1–11. https://www.scielo.org.ar/scielo.php?script=sci_arttext&pid=S2422-619X2021000200027
- Rezende, S. O., Marcacini, R. M., & Moura, M. F. (2011). O uso da mineração de textos para extração e organização não supervisionada de conhecimento. *Revista de Sistemas de Informação da FSMA*, 7, 7–21. <https://www.alice.cnptia.embrapa.br/alice/bitstream/doc/895476/1/FSMA.pdf>
- Silva, E. M. da, & Souza, R. R. (2014). Fundamentos em processamento de linguagem natural: Uma proposta para extração de bigramas. *Encontros Bibli: Revista Eletrônica de Biblioteconomia e Ciência da Informação*, 19(40), 1–32. <https://doi.org/10.5007/1518-2924.2014v19n40p1>
- Silva, S. R. de B., & Correa, R. F. (2020). Sistemas de indexação automática por atribuição: Uma análise comparativa. *Encontros Bibli: Revista Eletrônica de Biblioteconomia e Ciência da Informação*, 25, 1–15. <https://doi.org/10.5007/1518-2924.2020.e70740>
- Trucolo, C. C., & Digiampietri, L. A. (2014). Análise de tendências da produção científica nacional na área de Ciência da Informação: Estudo exploratório de mineração de textos. *AtoZ: Novas Práticas em Informação e Conhecimento*, 3(2), 87–94. <https://doi.org/10.5380/atoz.v3i2.41341>
- Wijffels, J. (2021). *textplot: Text plots* (Versão 0.2.1) [R package]. <https://CRAN.R-project.org/package=textplot>



Artículo disponible en acceso abierto bajo la Licencia CC Atribución-NoComercial-SinDerivadas.