
BibLex: aplicación práctica de IA y PLN para la asignación de palabras clave en registros de Biblat

BibLex: AI and NLP algorithm applied to keyword indexing into Biblat system

Patricia GARRIDO VILLEGAS, Edgar DURÁN MUÑOZ, Manuel Alejandro FLORES CHÁVEZ

Universidad Nacional Autónoma de México, Dirección General de Bibliotecas y Servicios Digitales de Información, Ciudad de México, México. pgarridov@dgb.unam.mx, eduranm@dgb.unam.mx, mafloresc@dgb.unam.mx

Resumen

Estudio aplicado de inteligencia artificial que describe la integración del algoritmo BibLex al proceso de indexación de artículos científicos, publicados en revistas latinoamericanas de acceso abierto indizadas en el portal Biblat. BibLex es un algoritmo diseñado para automatizar la asignación de palabras clave, combina técnicas clásicas de procesamiento de lenguaje natural (PLN), a través de la librería NLTK, con el modelo de inteligencia artificial generativa (GenIA) Gemini de Google. Ambos enfoques permiten analizar corpus lingüístico y semántico de los documentos, generando un conjunto de términos contextualizados. Los resultados sugieren que la integración de PLN con GenIA es técnicamente viable, asequible y oportuna para responder a los desafíos de la indexación bibliotecológica de manera eficiente y sin comprometer la calidad terminológica. El enfoque híbrido adoptado abre posibilidades para la automatización de procesos técnicos en los entornos bibliotecarios.

Palabras clave: Indexación automatizada. BibLex. Procesamiento de lenguaje natural (PLN). Inteligencia artificial generativa (GenIA). Biblat. CLASE. Periódica.

1. Introducción

El procesamiento de lenguaje natural (PLN en adelante) aplicado a la indexación facilita la identificación de términos que representan el contenido de un documento. Si bien, la labor del analista documental es necesaria para garantizar la adecuada elección de términos, el PLN puede optimizar el esfuerzo que el humano dedica a la extracción manual, así como a aumentar la exhaustividad y precisión de los términos de indexación, lo cual reduce la posibilidad de que un usuario pierda información relevante durante su proceso de búsqueda en un catálogo o sistema de información.

En el presente artículo se describe la integración y principales resultados del PLN en el proceso de indexación del portal Biblat, el cual indiza artículos provenientes de revistas científicas y académi-

Abstract

Applied artificial intelligence study describing the integration of the BibLex algorithm into the indexing process of scientific articles published in Latin American journals indexed by the Biblat portal. BibLex is an algorithm designed to automate keyword assignment by combining classical natural language processing (NLP) techniques, using the NLTK library, with Google's generative artificial intelligence model (GenIA), Gemini. Both approaches enable the analysis of the linguistic and semantic corpus of documents, generating a set of contextualized terms. BibLex is conceived as an effective solution to reduce manual indexing time, increase document processing capacity, and decrease terminological dispersion. The results suggest that integrating NLP with GenIA is technically feasible, cost-effective, and timely for addressing indexing challenges efficiently and without compromising terminological quality. The adopted hybrid approach opens possibilities automating technical processes in library environments.

Keywords: Automated indexing. BibLex. Natural language processing (NLP). Generative artificial intelligence (GenIA). Biblat. CLASE. Periódica.

cas de acceso abierto, editadas en América Latina y el Caribe. La aplicación del PLN se hizo a través de un algoritmo desarrollado *ex professo* y al que se denominó BibLex.

BibLex es un algoritmo en fase de pruebas para la extracción de palabras clave que reflejen el contenido de los documentos científicos indizados en el portal Biblat. Este desarrollo utiliza procesamiento de lenguaje natural e inteligencia artificial generativa para analizar semántica y sintácticamente textos completos de artículos y sugerir un conjunto de términos relevantes para describir el contenido. Fue desarrollado con la librería *Natural Language Toolkit* (NLTK), plataforma de código abierto que brinda interfaces y librerías de programación para trabajar con textos escritos en más de 50 idiomas. BibLex se encuentra implementado en el sistema de producción de registros

bibliográficos Biblat Central y sus resultados son validados por analistas documentales.

Aunado al algoritmo basado en técnicas clásicas del Procesamiento del Lenguaje Natural, se incorpora un modelo generativo avanzado (Gemini). La integración de Gemini en el proceso de indexación representa un complemento significativo al algoritmo BibLex, ya que permite profundizar en la interpretación semántica y en la generación contextualizada de términos clave. Gemini no se limita a detectar palabras frecuentes o explícitamente presentes en el texto, además de ello, analiza contextos implícitos, generando sugerencias conceptuales que enriquecen sustancialmente la cobertura temática de los artículos científicos y académicos.

Adicionalmente, aunque muchos modelos avanzados de inteligencia artificial generativa suelen requerir costos asociados al uso de una API, Gemini ofrece acceso gratuito dentro de ciertos límites mensuales. Esto facilita significativamente su integración en entornos académicos o institucionales como Biblat, al permitir que con su uso se validen y amplíen los términos propuestos sin incurrir en gastos adicionales, optimizando así los recursos y el tiempo destinado al análisis documental.

2. Antecedentes

El PLN es una rama de la inteligencia artificial dedicada al estudio y desarrollo de técnicas y herramientas computacionales capaces de analizar e interpretar el lenguaje humano (Bird et al., 2019; Kok et. al, 2009), une el conocimiento de 2 disciplinas, la computación y la lingüística para desarrollar técnicas que emulan las capacidades innatas del ser humano para entender y transmitir el lenguaje.

En el contexto de la propuesta, se entenderá por PLN cualquier tipo de manipulación informática del lenguaje natural para hacer comprensibles las expresiones humanas por las máquinas, hasta el punto de facilitar tareas intelectuales de los seres humanos (Klein et al., 2009, p. ix), por ejemplo, la indexación.

De acuerdo con Indurkha y Damerou (2010), en el PLN intervienen dos procesos:

- Análisis del Lenguaje Natural: conocido por sus siglas en inglés: NLA (Natural Language Analysis), es la parte del PLN que tiene como objetivo analizar un texto para llegar a extraer su significado de forma automatizada.
- Generación del Lenguaje Natural: en inglés NLG (Natural Language Generation). Son todas aquellas técnicas que tratan de generar

frases a partir de un conocimiento del sistema. No todos los sistemas que responden a preguntas son de tipo NLG, sino simplemente aquellos más avanzados que no devuelven una respuesta 'enlatada' y poco natural extraída de una base de datos de respuestas. La generan en tiempo real, partiendo de una información previa del sistema y de un modelo de generación del lenguaje aplicado a expresar dicha información.

Los algoritmos de PLN trabajan con módulos de funciones llamadas librerías. Se han hecho varios desarrollos de librerías que permiten realizar un análisis más preciso para el idioma español adaptándose a modelos de algoritmos ya existentes, así como a llevar a cabo el desarrollo de algoritmos adaptados a las características de los textos que se suelen analizar, esto último es el caso de esta propuesta.

Su objetivo, en el contexto de la bibliotecología, es realizar tareas con un mínimo de intervención humana, como la recuperación de información, clasificación documental, generación de respuestas orientadas a satisfacer las necesidades de información de los usuarios y, por supuesto, la indexación (Martínez, 2024).

2.1. Indexación por PLN

En entornos bibliotecológicos basados en colecciones digitales, el uso del PLN resulta fundamental para reducir las posibilidades de rezago bibliográfico y aumentar las posibilidades de éxito del usuario al buscar información (Miner, 2012). Un ejemplo de estos entornos son los sistemas de indexación de revistas de acceso abierto a nivel artículo, entre los que se pueden citar a Biblat. Este tipo de sistemas indexan documentos provenientes de revistas científicas y académicas previamente seleccionadas por criterios de calidad editorial. El problema radica en que la demanda de las revistas por ser indexadas aumenta y con ello, la cantidad de documentos que requieren un conjunto de términos para ser recuperables por el usuario.

La aplicación de PLN se basa en algoritmos que funcionan con ciertos modelos o representaciones del lenguaje. Los hay sencillos basados en la frecuencia con la que un término aparece en un documento, hasta más complejos, en donde intervienen modelos estadísticos para determinar el valor o peso de un término dentro del contexto del documento y determinar si incluirlo o no como descriptor. También es posible encontrar algoritmos que emplean ambas posibilidades, frecuencia y posición de las palabras. Entre los ejemplos utilizados para la asignación temática se encuentran la Asignación Latente de Dirichlet (ALD, en inglés

Latent Dirichlet Allocation), un modelo generativo propuesto por Blei, Ng y Jordan (2003). Según Polo Bautista & Martínez Acevedo (2021, p. 16)

La ALD es un modelo probabilístico generativo de un *corpus*, cuya idea básica es que los documentos se representan como mezclas aleatorias sobre temas latentes, donde cada tema se caracteriza por una distribución sobre palabras.

Uno de los objetivos principales de este algoritmo es optimizar el tiempo de identificación de temas de un documento, así como procesar una gran cantidad de información en menos tiempo. Dichas características son compartidas con el algoritmo propuesto en este trabajo.

Al igual, se ha desarrollado un algoritmo llamado RAKE (Rapid Automatic Keyword Extraction), que, según Contreras Barrera (2018, p. 115)

[...] es un algoritmo utilizado para la extracción de palabras clave —keywords— compuestas por una o más palabras, basado en las estadísticas de las palabras y de las coocurrencias de las mismas; trabaja sobre documentos individuales para obtener palabras clave compuestas por una o más de una palabra, las cuales sirven de base para la descripción del contenido de los documentos, la indización de los mismos o en algún estudio de minería de texto.

ALD también se utiliza para la clasificación temática. En función de la cercanía que tienen las palabras, el algoritmo establece un posible tema. Por ejemplo, las palabras biblioteca, fondo antiguo, digitalización, al aparecer juntas en diferentes partes del documento podrían generar el tema preservación digital. Por su parte, RAKE, analiza el texto y lo divide en partes, elimina conjunciones, preposiciones y otras palabras gramaticales para quedarse solo con la esencia del documento y buscar palabras o frases relevantes que se repiten y así construir los puntos de acceso temáticos. Este último algoritmo puede ofrecer resultados apropiados de entre el 25 y 60 %, según el estudio de Contreras (2018), mientras que ALD ofrece una precisión del 69 % comparado con términos propuestos por los autores de los documentos (Polo y Martínez, 2021).

Una de las limitantes del PLN para la asignación de palabras clave es que los modelos con los cuales funcionan los algoritmos tienden a entregar mejores resultados cuando se les entrena con un corpus propio, acotado a las necesidades del proceso de indización que se desea realizar (Cuéllar, 2025). Es por ello que el PLN tradicional puede ser complementado mediante otras tecnologías que permitan expandir las posibilidades de identificar términos relevantes, más allá de los enfoques basados en frecuencias o estructuras léxicas. En este sentido, la combinación de PLN con Inteligencia Artificial Generativa (GenIA, en

adelante) es un área de estudio que podría explorarse en la investigación.

2.2. Indización con GenIA

La GenIA se define como un área de la inteligencia artificial dedicada al estudio y desarrollo de técnicas computacionales para la creación de contenido basado en datos de entrenamiento, este contenido puede ser texto, imagen u otros medios (Sengar et al., 2024; Feuerriegel et al., 2023). Los ejemplos de GenIA incluyen aplicaciones como Chat GPT, DALL-E o Copilot. Con base en Gatti (2024), puede inferirse que una de las diferencias clave entre la GenIA y el PLN radica en que la primera cada vez requiere menos instrucciones programadas y busca operar a partir de una instrucción escrita en lenguaje cotidiano, llamada *prompt*.

La eficacia de los modelos de GenIA para efectuar tareas de indización dependerá de sus capacidades técnicas. Bouzid y Piron (2024) estudiaron las posibilidades de ChatGPT, Claude y Gemini para crear términos de indización a partir de un conjunto de resúmenes de documentos científicos. ChatGPT fue el modelo que más términos entregaba, entre 20 y 25, llegando a tener similitud del 96 % en términos otorgados por Claude. Los autores antes citados también encontraron que a medida que se solicitaban más términos, los modelos generaban meros sinónimos y términos derivados de resultados previos. En cuanto a precisión, ChatGPT fue más preciso y exhaustivo y Gemini tuvo resultados más modestos.

En otro estudio, el uso de los modelos Donut y GPT facilitó la creación de índices para la documentación técnica en la industria de la construcción, con una precisión de 85 % para Donut y del 86% en el caso de GPT (Feyisa et al., 2024). Los autores del estudio entrenaron a estos modelos con el trabajo manual previo de términos extraídos de la documentación técnica, almacenados en formato JSON y en bases SQL, de tal manera que los modelos fueran capaces de entregar como resultado el conjunto de títulos, subtítulos y su paginación para conformar los índices y así reducir el grado de intervención del recurso humano.

La indización por PLN o por GenIA ha sido estudiada por separado, lo cual abre una brecha para indagar en las posibilidades de integrar ambos tipos de inteligencia artificial dentro de un proceso de obtención de palabras clave. Mientras que el PLN clásico facilita la extracción precisa y estructurada de términos clave mediante reglas y análisis lingüísticos específicos, la GenIA nos permite añadir valor al proporcionar respuestas flexibles y contextualizadas.

3. Problema y metodología

Biblat es un portal de revistas científicas de acceso abierto editadas en América Latina y el Caribe, ofrece información en texto completo de más de 350,000 documentos, entre los que se encuentran artículos, reseñas de libros, cartas editoriales, entre otros. Además, cuenta con un servicio de localización de documentos que por su antigüedad solo se encuentran en revistas de formato impreso y que se tienen en el acervo de la Hemeroteca Latinoamericana de la UNAM. La principal fuente de información de Biblat son dos bases de datos fundadas hace ya más de 40 años, CLASE (Citas Latinoamericanas en Ciencias Sociales y Humanidades) y Periódica (Índice de Revistas Latinoamericanas en Ciencias).

La indización de documentos en Biblat se basa en las convenciones bibliotecológicas que, en lo general, apuntan a hacer una extracción de los términos temáticos, onomásticos y geográficos que representan con la mayor fidelidad posible el contenido de un documento y facilitan que el usuario pueda recuperarlo. Reyna (2015) señaló que la experiencia y las particularidades del tipo de documentos que se indizan en Biblat ha llevado a la redacción de un manual propio, el cual, establece los pasos y los criterios que el analista debe considerar.

El método de indización en Biblat ha sido en su mayor parte manual, esto quiere decir que los analistas a partir de la lectura del título, resumen y partes clave del documento determinan cuáles términos utilizar para representar el contenido. Los términos son ingresados en una plantilla de Aleph y comprobados con el listado autorizado que tiene cargo el sistema (Alonso, Arana, Reyna y Sánchez, 2012).

La cantidad de documentos que integran Biblat ha aumentado con los años, lo que ha provocado un atraso en la actualización e indización de dicha información, aunado a la necesidad de más de personal académico para el análisis documental en el Departamento de Bibliografía Latinoamericana, es por esa razón que un algoritmo de PLN e inteligencia artificial generativa sería de gran ayuda.

Biblat requiere procedimientos que faciliten el control y actualización de su listado de palabras clave, a través del algoritmo, dicha tarea será más sencilla y de manera indirecta ayudará a elevar la producción de los analistas debido a que la asignación de palabras clave es uno de los procesos intelectualmente más complejos. El algoritmo de PLN e IA Generativa que se propone en este proyecto se dirige a apoyar al analista documental para los campos de *Palabras clave* y *Keywords*.

Es importante mencionar que, si bien muchos modelos avanzados de inteligencia artificial generativa requieren el pago por el uso de sus APIs, en este procedimiento se opta por Gemini debido a que ofrece un nivel de acceso sin costo, dentro de ciertos límites mensuales. Según la documentación oficial de Google (2025), la API de Gemini establece restricciones específicas para este nivel gratuito. Esto facilita notablemente su integración en entornos académicos o institucionales como Biblat, al permitir la validación y ampliación de los términos propuestos sin incurrir en gastos adicionales significativos, optimizando así los recursos y el tiempo destinado al análisis documental.

Los límites de uso de la API se estructuran en tres dimensiones clave: solicitudes por minuto (RPM), solicitudes por día (RPD) y tokens por minuto (TPM) en la entrada. El sistema evalúa el uso en función de cada uno de estos parámetros, y si se excede alguno, se genera un error de límite de frecuencia, independientemente de si los otros parámetros permanecen dentro de los márgenes permitidos. Por ejemplo, si un proyecto tiene un límite de 20 RPM y se realizan 21 solicitudes en un solo minuto, se bloqueará temporalmente el acceso, incluso si el consumo de tokens se encuentra dentro del límite establecido.

En el presente estudio se emplea la API de Gemini 2.0 Flash. Este modelo opera bajo un esquema de uso gratuito que establece límites de 15 solicitudes por minuto (RPM), 200 solicitudes por día (RPD) y hasta 1,000,000 tokens de entrada por minuto (TPM). Estas cuotas permiten procesar, en condiciones óptimas, hasta 15 artículos por minuto y un total de 200 artículos diarios, siempre que cada uno se envíe en una sola solicitud. Considerando un promedio de 1,500 tokens por página, el límite de tokens por minuto habilita el análisis de aproximadamente 600 páginas de texto académico por minuto, lo cual es suficiente para trabajar con artículos de entre 20 y 30 páginas sin exceder el umbral permitido. Esta capacidad es adecuada para tareas de procesamiento masivo de documentos dentro de flujos de trabajo de indización o análisis de contenido científico.

4. Integración de BibLex en la metodología de indización en Biblat

Se expone la incorporación al proceso de análisis documental de BibLex a través de un algoritmo de Procesamiento de Lenguaje Natural junto con un algoritmo de IA Generativa para la indización automatizada de palabras clave de los artículos de las revistas académicas indizadas en el portal Biblat. La principal diferencia entre la indización manual y la realizada mediante algoritmos radica

en la metodología aplicada: mientras que la indicación manual se basa en el juicio experto y en la lectura interpretativa del contenido por parte de especialistas (Figura 1), la indicación automatizada opera mediante modelos computacionales que extraen, procesan y asignan términos clave a partir de patrones lingüísticos y estadísticos. Esta distinción metodológica no sólo refleja un cambio en la forma en que se realiza la indicación, sino que también implica una mejora en términos de eficiencia, escalabilidad y consistencia.

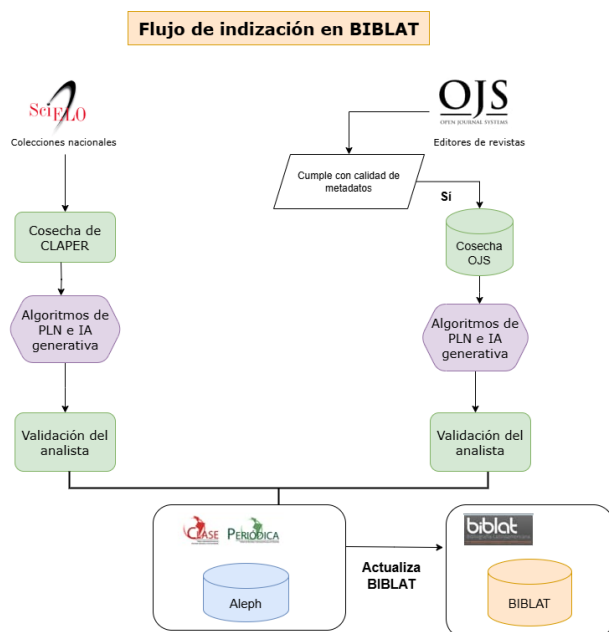


Figura 1. Flujo general de indización en Biblat con BibLex

En la Figura 2 se muestra un esquema que resume las fases del proceso de indización automatizada de las palabras clave.

El punto de partida es la obtención del texto de los documentos, para que el texto de cada artículo sea procesable por los algoritmos empleando PLN y GenIA, se extrae de los archivos PDF utilizando el lenguaje de programación Python y sus bibliotecas PyPDF2 y pdfminer. El flujo básico es: se carga del PDF en memoria y se realiza la conversión de cada página a texto plano; se eliminan saltos de línea erráticos, símbolos de control y duplicados de espacios; se convierten caracteres a UTF-8 y se corrigen guiones de fin de renglón.

Respecto al uso de Gemini, una vez obtenido el corpus limpio, se construye un *prompt* que concatena, en primer lugar, una instrucción breve en la que se solicita la generación de palabras clave, incluyendo algunas reglas básicas utilizadas por el Departamento de Bibliografía Latinoamericana; en segundo lugar, se incorpora el bloque textual

extraído del documento. La llamada se hace desde la biblioteca `google.generativeai`, donde se inicializa el modelo `gemini-2.0-flash` y se envía la solicitud vía el método `generate_content()`. La respuesta es obtenida en formato JSON lista para su almacenamiento en Biblat Central.

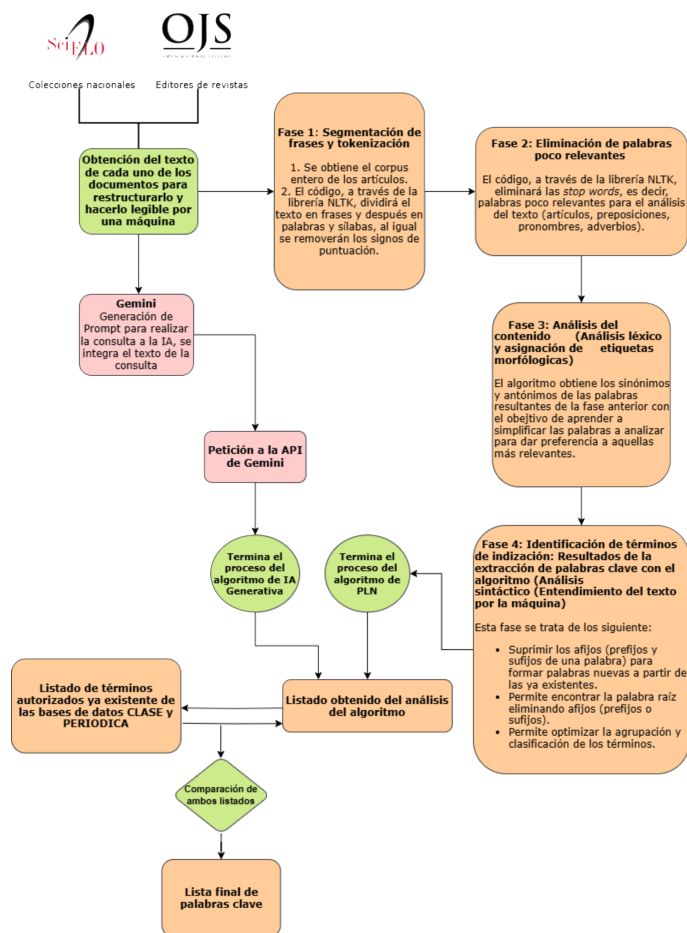


Figura 2. Fase de indización automatizada con BibLex

Desde la perspectiva de la interfaz web empleada para la generación de palabras clave, a continuación, se describe, a grandes rasgos, el flujo de trabajo en el sistema Biblat Central.

El sistema contempla una funcionalidad específica para la asignación de registros a los analistas. Dentro de este módulo, se realiza también la generación automatizada de palabras clave previa a dicha asignación.

El módulo de asignación, como se muestra en la Figura, presenta un listado que incluye todos los fascículos cosechados, provenientes tanto de colecciones SciELO como de revistas alojadas en OJS. La primera etapa del proceso consiste en una revisión de ciertos metadatos descriptivos, tales como el título del artículo, idioma, disciplina, autores y sus instituciones de afiliación.

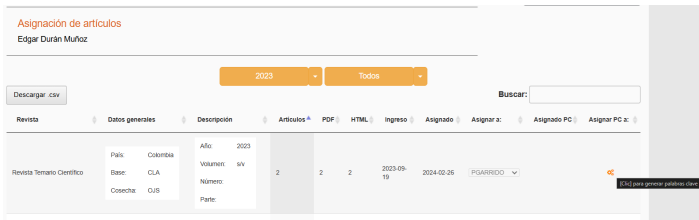


Figura 3. Interfaz para generación de palabras y asignación de fascículos

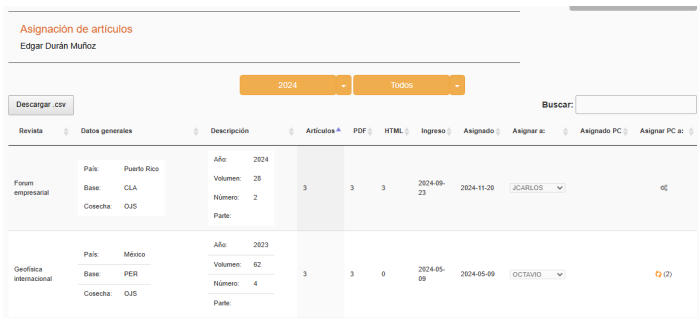


Figura 4. Generación en segundo plano de palabras clave

Una vez concluida esta primera revisión, en la columna “Asignar a:” del listado se visualiza el nombre del analista responsable, mostrado de forma

deshabilitada dentro de un componente de selección. Simultáneamente, en la columna “Asignar PC a:”, se despliega un ícono de engranes que indica que los artículos se encuentran listos para iniciar el proceso automático de generación de palabras clave (Figura 3).

Para iniciar dicho proceso, basta con hacer clic sobre el ícono correspondiente. El sistema ejecutará esta tarea en segundo plano. El tiempo requerido para completarla dependerá del volumen de documentos a procesar, de la velocidad de comunicación con la página de la revista, así como del rendimiento de la API de Gemini y del motor de procesamiento PLN. Durante la ejecución, el sistema actualiza el listado indicando el número de artículos que han sido procesados (Figura 4).

Al finalizar, se vuelve a mostrar el listado de analistas disponibles, lo cual indica que el fascículo está listo para ser asignado a la etapa de revisión de palabras clave. Una vez realizada esta asignación, el analista encontrará en su lista de trabajo los artículos correspondientes. Al seleccionar uno de ellos, se mostrarán las palabras clave generadas tanto por el sistema de PLN como por la inteligencia artificial generativa Gemini (Figura 5).

A continuación se muestran las palabras claves sugeridas, palabras asentadas por autores y palabras extraídas del texto.

(I) Selección las adecuadas para el artículo.

(II) Edite la palabra si determina que existe un término más adecuado para sustituir, considere que en adelante si se encuentra nuevamente el mismo término, se realizará la misma sustitución de manera automática.



Figura 5. Interfaz para la revisión de palabras clave generadas automáticamente

A la derecha de cada término se indica la cantidad de ocurrencias registradas en la base de datos. Esta información tiene como objetivo orientar

al analista sobre cuáles forman parte del catálogo y cuáles son los más utilizados.

Los resultados muestran que una proporción significativa coincide con registros existentes, lo que sugiere un buen desempeño del algoritmo en la asignación. Sin embargo, la selección final de los términos aún queda bajo el criterio del analista.

Al utilizar modelos de inteligencia artificial para la extracción automática de palabras clave a partir de textos completos de artículos, es importante considerar las implicaciones legales y éticas relacionadas con las licencias Creative Commons, particularmente aquellas que contienen cláusulas No Comercial (NC) y Sin Derivadas (ND). Según lo expuesto por Creative Commons (2025), los contenidos con licencia ND no pueden utilizarse para entrenar modelos de IA, ya que se considera una forma de creación derivada, y los contenidos NC no pueden emplearse con fines comerciales, lo que incluye ciertas plataformas gratuitas de IA cuyo uso implica la posibilidad de que los datos ingresados se utilicen para futuros entrenamientos comerciales.

En nuestro caso, aunque el acceso a herramientas como Gemini se realiza de forma gratuita, esto podría vincular indirectamente el proceso con fines comerciales no deseados. No obstante, nuestro trabajo con textos disponibles en plataformas como la base de datos Biblat se realiza con fines estrictamente académicos, y sin ánimo de lucro y con vistas a generar descriptores de acceso a una obra, no a la reproducción de ésta. El propósito es facilitar la visibilidad, el acceso y la recuperación de conocimiento científico mediante el uso de tecnologías accesibles, sin redistribuir ni modificar los contenidos originales ni desarrollar herramientas comerciales sobre ellos. Aun así, reconocemos que estas prácticas se sitúan en un área legal y ética aún en desarrollo, lo que refuerza la necesidad de actuar con responsabilidad y transparencia.

Una posible vía para evitar esta zona gris legal sería el uso de modelos de lenguaje de código abierto (LLM) ejecutados en una infraestructura propia, lo cual permitiría un mayor control sobre el procesamiento de datos sin involucrar plataformas comerciales. Sin embargo, esta alternativa implica costos adicionales en términos de recursos computacionales, infraestructura y mantenimiento, que pueden ser poco viables para dependencias pequeñas con recursos limitados.

5. Conclusiones

La incorporación de BibLex al flujo de trabajo de indización del portal Biblat representa un avance significativo en la automatización de procesos documentales dentro del ámbito bibliotecológico. Este desarrollo, basado en un algoritmo propio de Procesamiento de Lenguaje Natural (PLN)

complementado con un modelo de inteligencia artificial generativa (Gemini), permite una extracción precisa y contextualizada de palabras clave a partir del texto completo de los documentos, contribuyendo a una mejora sustantiva en la calidad de los registros bibliográficos.

El sistema propuesto presenta una metodología robusta, centrada en la identificación de términos relevantes más allá de la frecuencia de aparición, y considerando estructuras gramaticales y contextuales en español e inglés. Esta metodología ha sido integrada exitosamente al sistema de producción de registros bibliográficos Biblat Central, manteniendo una estrecha alineación con el listado de términos autorizados construido históricamente por los analistas documentales.

Entre los principales beneficios observados con la implementación de BibLex, destacan los siguientes:

- Uno de los aportes más relevantes de BibLex radica en el cambio metodológico que introduce en el proceso de indización: a diferencia del enfoque manual, que se basa en la lectura interpretativa y el juicio experto de los documentalistas, la indización automática utiliza técnicas computacionales para analizar el contenido textual, identificar patrones semánticos y sugerir términos clave con base en criterios lingüísticos y estadísticos, reduciendo la carga cognitiva que conlleva la lectura humana. Esta diferencia metodológica no implica una sustitución del conocimiento experto, sino una transformación del proceso, en el cual la automatización puede actuar como una etapa preliminar que potencia la calidad, consistencia y representatividad temática del análisis documental.
- Simplificación de la asignación temática, dado que los analistas tendrán la posibilidad de seleccionar términos clave basados en la relevancia sin dedicar demasiada atención a la revisión del contenido. La asignación temática es quizá uno de los pasos que más atención y tiempo demanda en el análisis documental, pero con ayuda del algoritmo, en el futuro, podría verse un beneficio.
- Alineación terminológica con los estándares de calidad de Biblat, gracias a que el algoritmo se diseñó para operar en función de los vocabularios controlados existentes, preservando así la consistencia semántica del sistema.
- Bajo costo de implementación, al utilizar la API gratuita de Gemini no se requieren hardware adicional, licencias de software ni personal especializado para la operación cotidiana. No obstante, esta condición depende de que

se mantengan los límites de uso gratuito de Gemini y las licencias Creative Commons asociadas a los artículos procesados. Si estos parámetros se volvieran más restrictivos, podría ser necesario migrar a un modelo abierto con características equiparables o que impacten lo menos posible en la funcionalidad del algoritmo. En ese escenario, la adopción del nuevo modelo implicaría invertir en infraestructura para cubrir los requisitos de hardware correspondientes.

- Facilidad de adopción institucional, al integrarse sin fricciones en los procesos internos del Departamento de Bibliografía Latinoamericana y sin necesidad de capacitación tecnológica compleja para los analistas.

Finalmente, BibLex evidencia el potencial de la convergencia entre técnicas tradicionales de PLN e innovaciones recientes en IA generativa para resolver desafíos históricos en los sistemas de recuperación de información. Su adopción no sólo fortalece la eficiencia operativa, sino que también habilita nuevas posibilidades en la elaboración de productos bibliométricos y en la mejora continua de la visibilidad, accesibilidad y análisis temático de la producción científica regional.

Declaración de autoría

Patricia Garrido Villegas: Conceptualization (lead); Software (equal); Investigation (lead); Validation (equal); writing – original draft (equal); formal analysis (equal); writing – review and editing (equal); Methodology (equal).

Edgar Durán Muñoz: Software (equal); Visualization (lead); Investigation (lead); Validation (equal); writing – review and editing (equal); review and editing (equal); formal analysis (equal); Methodology (equal).

Manuel Alejandro Flores Chávez: Methodology (equal); writing – review and editing (equal). Conceptualization (supporting); Writing – original draft (supporting); Writing – review and editing (lead); formal analysis (equal); Investigation (supporting).

Agradecimientos

Los autores expresan su agradecimiento a los analistas Lic. Marco A. Flores Montes, Lic. Nidia Zúñiga Murrieta, Lic. Juan Carlos Díaz Mauricio y Lic. Flor Janet Rivera Pulido por las pruebas realizadas en la plataforma de asignación automática de palabras clave. Asimismo, se agradece a la Mtra. María Guadalupe Trinidad Argüello Mendoza por su apoyo en la gestión y administración de la asignación de documentos.

Referencias

- Alonso Gamboa, José Octavio; Arana Mendoza, Celia; Reyna Espinosa, Felipe Rafael; Sánchez Pereyra, Antonio (2012). Manual de indexación para las bases de datos Clase y Periódica. México, D.F.: Universidad Nacional Autónoma de México, Dirección General de Bibliotecas. https://biblat.unam.mx/archivos/manual_indexacion.pdf
- Bird, Steven; Klein, Ewan y Loper Edward (2009). Natural Language Processing with Python. O'Reilly.

- Bouzid, Sara y Piron, Loïs (2024). Leveraging Generative AI in Short Document Indexing. // *Electronics*. 13:17 3563. <https://doi.org/10.3390/electronics13173563>
- Blei, D. M.; Ng, A. Y.; Jordan, M. I. (2003). Latent Dirichlet Allocation. // *Journal of Machine Learning Research*, 3(Jan), 993–1022. <https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf>
- Contreras Barrera, Marcial (2018). Aplicación del algoritmo RAKE en la indexación de documentos digitales. // *Investigación Bibliotecológica*. 32:75 (abril/mayo) 109-123. <https://doi.org/10.22201/iibi.24488321xe.2018.75.57951>
- Creative Commons (2025). Using CC-Licensed Works for AI Training. <https://creativecommons.org/using-cc-licensed-works-for-ai-training-2>
- Cuéllar Hidalgo, Rodrigo (2025). Reconocimiento de Entidades Nombradas (NER): una técnica para agilizar el procesamiento de colecciones digitales. // *Amontonamos las palabras: Blog de la Biblioteca de El Colegio de México*. <https://doi.org/10.58079/13dqq>
- Feuerriegel, Stefan; Hartmann, Jochen; Janiesch, Christian y Zschech, Patrick (2023). Generative AI. // *arXiv*. <https://doi.org/10.48550/arXiv.2309.07930>
- Feyisa, Degaga Wolde; Berihun, Haylemicheal; Zewdu, Ammanuel; Najimoghadam, Mahsa y Zare, Marzieh (2024). The Future of Document Indexing: GPT and Donut Revolutionize Table of Content Processing. // *arXiv*. <https://doi.org/10.48550/arXiv.2403.07553>
- Gatti, Alberto (2024). Alfabetización e inteligencia artificial. // *Journal of Neuroeducation*. 5:1 (julio) 52–58. <https://doi.org/10.1344/joned.v5i1.46108>
- Google (2025). Límites de frecuencia. // *Google AI for Developers*. <https://ai.google.dev/gemini-api/docs/rate-limits?hl=es-419>
- Indurkhya, N.; Damerau, F. J. (Eds.). (2010). Handbook of Natural Language Processing. Chapman and Hall/CRC. <https://doi.org/10.1201/9781420085938>
- Kok, Joost N.; Boers, Egbert. J. W.; Kusters, Walter A.; Van der Putten, Peter; Poel, Mannes (2009). Artificial intelligence: definition, trends, techniques, and cases. // *Artificial intelligence*. 1, 270-299.
- Klein, E.; Bird, S.; Loper, E. (2009). Natural Language Processing with Python. O'Reilly Media, Incorporated.
- Martínez Albarrán, Alí (2024). La inteligencia artificial en los estudios de la información y la bibliotecología. // *e-Ciencias de la información*. 14:2. <https://doi.org/10.15517/eci.v14i2.57949>
- Miner, Gary; Delen, Dursun; Elder, John; Thomas Hill; Nisbet, Robert A. (Eds) (2012). Practical Text Mining and Statistical Analysis for Non-Structured Text Data Applications. // Elsevier Science & Technology. <https://doi.org/10.1016/C2010-0-66188-8>
- Polo Bautista, Luis Roberto y Martínez Acevedo, Karen Vanessa (2021). Algoritmo para el análisis temático de documentos digitales. // *Investigación Bibliotecológica*. 35:89 (octubre/diciembre) 13-31. <https://doi.org/10.22201/iibi.24488321xe.2021.89.58419>
- Reyna Espinosa, Felipe Rafael (2015). CLASE. Perfil de una base de datos bibliográfica. // *Biblioteca Universitaria* 18:2. https://biblat.unam.mx/hevila/e-BIBLAT/Biblio/ReynaEspinosa_2015%282%29.pdf
- Sengar, Sandeep Singh; Hasan, Affan Bin; Kumar, Sanhay y Carroll, Fiona (2024). Generative Artificial Intelligence: A Systematic Review and Applications. // *arXiv*. <https://doi.org/10.48550/arXiv.2405>

Enviado: 2025-03-31. Segunda versión: 2025-09-24.
Aceptado: 2025-10-17.