eib0985883306

Identificación de las temáticas de investigación del Chocó en la literatura indizada en Scopus

Cristina Restrepo-Arango*

Artículo recibido:
3 de agosto de 2023
Artículo aceptado:
13 de diciembre de 2023

Artículo de investigación

RESUMEN

El objetivo de esta artículo radica en extraer las temáticas de investigación de los resúmenes y datos bibliográficos de los artículos indexados en la base de datos Scopus y que tienen como objeto de estudio al departamento del Chocó (Colombia). De esta manera, se buscaron las palabras clave Chocó AND Colombia en la base de datos Scopus, se exportaron las referencias bibliográficas a EndNote y se extrajeron los datos de autor(es), título, publicación periódica, volumen, número, año y resumen, se convirtieron en un archivo de texto, se eliminaron referencias y símbolos. La manipulación del archivo en pdf se realizó con la ejecución de preparación del texto, *tokenización*, lematización y obtención de lista de *bigrams*

* Oficina de Bibliotecas y Recursos Educativos, Universidad de Córdoba, Colombia.
crestrepoarango@gmail.com

que se efectuaron en el entorno de desarrollo integrado (EDI) de RStudio. Así, se encontraron 668 registros bibliográficos de documentos indexados en Scopus. Las palabras con el mayor número de frecuencia de aparición: «species», «Colombia», «Chocó», «forest», «pacific», «tropical», etcétera. Se encontraron 89 841 *bigrams*, entre los que destacan «new species», «pacific coast», «colombian pacific», entre otros. Las colocaciones de palabras muestran que «gold» combina con «mining», «mercury», «platinum», y así sucesivamente. «Chocó» combina con «Colombia», «biogeographical», «rain», «tropical», y demás. «Biodiversity» combina con «conservation», «tropical», «agricultural», etcétera. «Climate» combina con «change», «variability», «basin», y más. Se concluye que las palabras más frecuentes evidencian que hay una preocupación por el estudio de la minería, la biodiversidad, el cambio climático, el bosque tropical, el océano pacífico, entre otros.

Palabras clave: Minería de textos; Colocaciones; Co-ocurrencia de palabras; Chocó (Colombia)

Identification of the research themes of Chocó in the literature indexed in Scopus

Cristina Restrepo-Arango

ABSTRACT

Objective: extract the research topics from the summaries and bibliographic data of the articles indexed in the Scopus database that have the department of Chocó (Colombia) as their object of study. Methods: The keywords Chocó AND Colombia were searched in the Scopus database, the bibliographic references were exported to EndNote and the data of author(s), title, periodical publication, volume, number, year and abstract were extracted, they were converted into a text file, references and symbols were removed. The manipulation of the pdf file was carried out with the execution of text preparation, tokenization, lemmatization and obtaining the list of bigrams that were carried out in the integrated development environment (EDI) of RStudio. Results: 668 bibliographic records of indexed documents were found in Scopus. The words with the highest frequency of occurrence are: «species», «Colombia», «Chocó», «forest», «pacific», «tropical», etc. 89 841 bigrams were found, including «new species», «pacific coast», «colombian pacific», etc. Word collocations show that «gold» matches «mining»,

«mercury», «platinum», etc. «Chocó» combines with «Colombia», «biogeographical», «rain», «tropical», etc. «Biodiversity» combines with «conservation», «tropical», «agricultural», etc. «Climate» combines with «change», «variability», «basin», etc. Conclusions: the most frequent words show that there is a concern for the study of mining, biodiversity, climate change, the tropical forest, the Pacific Ocean, etc.

Keywords: Text mining; Collocations; Co-occurrence of words; Choco (Colombia).

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2024.98.58833

Introducción

El origen de la minería de textos se encuentra en las áreas de investigación de bases de datos, aprendizaje automático y estadística. Las bases de datos se requieren para almacenar, acceder y analizar grandes cantidades de información. El aprendizaje automático (*machine learning*) representa un área de la inteligencia artificial relacionada con el desarrollo de técnicas que permiten a los computadores aprender, por medio del análisis de conjuntos de datos. La estadística tiene sus bases en las matemáticas y se ocupa de la ciencia y la práctica para el análisis de datos empíricos (Hotho, Nürnberger y Paaß 2005). El uso de la minería de textos para extraer información –a partir de los documentos producidos y publicados por los científicos– ha tomado fuerza con la aparición de la inteligencia artificial. Su propósito primordial radica en identificar relaciones e interacciones temáticas entre conceptos. Las temáticas identificadas proporcionan a los científicos ideas concretas para explorar nuevos campos, o bien, fortalecer los campos de investigación existentes. La minería de textos tiene como objeto de estudio la producción académica que se publica en un área del conocimiento, o bien, los datos y las informaciones que se comparten en las redes sociales. Esto significa que la información debe estar en un soporte que sea legible por máquinas, pues, la minería de textos aparece gracias a los desarrollos de la computación.

La minería de textos o descubrimiento de conocimiento a partir del texto (*knowledge discovery from text*, en sus siglas en inglés KDT), tiene como propósito «revelar la información oculta, por medio de métodos que, […] son capaces de hacer frente a la gran cantidad de palabras y estructuras en el lenguaje natural [que] permite manejar la vaguedad, la incertidumbre y la borrosidad» (Hotho, Nürnberger y Paaß 2005, 2). La minería de textos usa métodos que permiten explorar la información textual que se publica en Internet en diferentes formatos, sobre todo ayuda a la exploración de patrones que se encuentran en los datos

no estructurados. Estos constituyen imágenes, audios, datos de texto, datos cartográficos, etcétera. Normalmente tienen una estructura interna, son generados por el ser humano o una máquina en formato textual o no textual. En este caso los artículos publicados en revistas científicas se consideran datos no estructurados, porque su elaboración, apartados y formalismos de forma y de fondo responden a las indicaciones propias de cada revista. Es así como la minería de textos ayuda a identificar patrones relacionados con las palabras contenidas en los resúmenes, títulos o desarrollo de los artículos.

La minería de textos posibilita que se puedan «descubrir tendencias, patrones, desviaciones y asociaciones de una colección de textos […] en considerables cantidades de información no estructurada» (Contreras 2014, 131) como es el caso de la información académica, la cual se caracteriza por tres aspectos. Primero, presenta resultados de investigación contenidos en datos textuales, datos numéricos, imágenes, etcétera. Segundo, la información es publicada en canales de comunicación formales o informales (redes sociales). Tercero, la información académica no está estructurada; por ejemplo, libros, informes de investigación, etcétera, y el investigador utiliza un vocabulario técnico para referirse al fenómeno o problema de investigación.

La minería de textos se complementa con el procesamiento del lenguaje natural (PLN) para extraer información significativa, por medio de paqueterías o softwares especializados que aplican procesos de lematización, fragmentación y derivación para la identificación de «n-gramas» y «tokens», entre otras formas que aparecen en un conjunto de documentos. Además, ambas técnicas usan la co-ocurrencia de las palabras para captar semántica y sintácticamente las relaciones entre palabras, con el fin de que los computadores entiendan y comprendan de la mejor manera el lenguaje humano (Trask, Gilmore y Russell 2015; Russell 2013). También, las técnicas de PLN y minería de textos resultan ampliamente usadas para analizar los sentimientos contenidos en las publicaciones (Alkan, Karakuş y Direkci 2023; Ma *et al.* 2023).

En general, estas técnicas se usan para analizar el lenguaje humano a partir de la eliminación «stopwords», es decir, de la eliminación de artículos, preposiciones, conjunciones, signos de puntuación, etcétera. Después de este proceso se realiza la derivación (*stemming*) que convierte las palabras a su forma simple. Por último, la fragmentación consiste en convertir un texto en oraciones y luego en «tokens» o palabras. La tokenización de oraciones puede presentar problemas al romper una oración por contener palabras abreviadas en un texto, las cuales normalmente llevan al final un punto (Russell 2013). Este tipo de fallas en la tokenización las ejemplifica Russell (2013) con oraciones como «Mr. Green killed Colonel Mustard in the study with the candlestick». Seguramente el tokenizador de oraciones no extraerá a Mr. Green en la tokenización.

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2024.98.58833

La aplicación de estos procesos permite obtener datos que facilitan el análisis de co-ocurrencias de palabras (ACP) y el análisis de colocaciones.

El ACP permite mapear la estructura intelectual de un dominio específico, por medio del conteo y análisis de las co-ocurrencias de palabras contenidas en unidades bibliográficas como artículos académicos, ponencias, capítulos de libros, entre otros. Básicamente el análisis de co-ocurrencias explora la red conceptual de las palabras clave para identificar relaciones temáticas en diferentes campos del conocimiento. Se considera un método cuantitativo para mapear las relaciones e interrelaciones entre conceptos, ideas y problemas en diferentes campos científicos (Hosseini *et al*. 2021). Este tipo de análisis se fundamenta en que «dos o más palabras se relacionan entre sí en un significado semántico, [en tanto] si coocurren en el mismo documento» (Yin 2020, 1886). El análisis de co-ocurrencia de palabras permite descubrir conceptos y sus relaciones en un documento.

El análisis de colocaciones se sustenta en que ninguna palabra aislada tiene significado por sí sola. Según Firth (1957, 6), «conoces a una palabra por sus compañeras», esto es, que el significado de una palabra se basa en el significado de las palabras que la rodean. Para Corpas Pastor (2001), las colocaciones representan unidades fraseológicas que están formadas por al menos dos palabras; por ejemplo, «cambio climático», «minería artesanal», etcétera, de no ser así, por palabras que si bien no están unidas tienen una relación sintáctica o de significado. Esas unidades fraseológicas constituyen grupos de palabras que se convierten en términos estandarizados por el uso que hacen de dichos grupos de palabras una comunidad académica. También, se puede entender por colocación combinaciones de palabras que resultan teóricamente posibles y consumadas por los hablantes, o bien, para este caso los autores de los artículos académicos publicados en Scopus sobre esta temática.

Ambas técnicas se complementan y se basan en las palabras clave contenidas en los títulos, los resúmenes y los descriptores que forman parte inherente de los textos producidos por los científicos. Por ello, con la aplicación de las técnicas usadas en la minería de textos se tiene como propósito extraer las temáticas de investigación de los resúmenes y datos bibliográficos de los artículos indexados en la base de datos Scopus que tienen como objeto de estudio al departamento del Chocó (Colombia). Para ello se intentará dar respuesta a las siguientes preguntas:

> ¿Cómo puede contribuir el procesamiento del lenguaje natural a la organización de la información y al conocimiento de la bibliotecología?
> ¿Cuáles son las palabras con mayores frecuencias de aparición?
> ¿Cuáles son los pares de palabras o *bigrams* que aparecen en el corpus?
> ¿Cuáles son las interacciones entre las palabras o la co-ocurrencia de palabras?

¿Cuáles son las relaciones temáticas o colocaciones de las palabras «gold», «Chocó», «biodiversity», «climate» y «Atrato»?

El departamento del Chocó conforma un departamento colombiano que se localiza en el occidente, limita con la República de Panamá y el mar caribe. Su economía depende de la minería que se concentra en la extracción de oro, plata y platino, también de la explotación forestal, así como de la agricultura y la ganadería. Integra uno de los departamentos con una gran extensión de selva y de alta pluviosidad, tiene costas sobre los océanos Atlántico y Pacífico. Además, cuenta con dos importantes ríos, el Atrato y San Juan, que constituyen las principales vías de transporte. La mayor parte de la población está compuesta por: afrodescendiente (75,68 %), le siguen indígenas (11,9 %), mestizos (7,42%) y blancos (5,01 %) (Gobernación del Chocó 2023). Las características geográficas de este departamento lo convierten en un objeto de investigación importante para los científicos de diferentes disciplinas, más aún, al no encontrar estudios similares aplicando esta técnica.

Destaca que, con la aplicación de la minería de textos, no sólo se extraen tendencias en investigación sobre el departamento del Chocó en los documentos indexados en Scopus, sino que además se están aportando a la bibliotecología y a la documentación técnicas que permiten identificar conceptos, lugares geográficos, nombres, etcétera, los cuales se usan en la literatura científica y se pueden utilizar en la organización de la información para representar la literatura existente en un campo del conocimiento, y de esta manera facilitar la recuperación de la misma. Por igual, dichas técnicas se pueden usar para la construcción de ontologías y tesauros.

Revisión de literatura

La revisión de literatura se centró en identificar tres tipos de estudios publicados. Primero, artículos que hubieran aplicado la minería de textos en la bibliotecología y documentación; segundo, artículos que usaron el lenguaje de programación R en la bibliotecología y documentación; tercero, estudios que aplicaron las técnicas de minería de textos, el PLN y el lenguaje de programación R.

Se identificaron artículos que aplicaron la minería de textos en la bibliotecología y la documentación, como el caso del trabajo de Al-Betar *et al.* (2023), quienes aplican la agrupación de documentos en función de la similitud de su contenido para extraer palabras clave de los artículos. También Zhang *et al.* (2023) proponen un modelo para extraer las palabras clave basado en un modelo de gráfico jerárquico semántico. El método propuesto tiene en cuenta el contexto

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2024.98.58833

interno y la relación que establecen las palabras clave. Shen (2023) usó SciBERT para mejorar las tareas que realiza el procesamiento del lenguaje natural y aplicó un modelo previamente entrenado basado en los resúmenes publicados en las revistas *Social Science Citation Index* (SSCI).

Asimismo, la revisión de la literatura identificó estudios publicados que utilizaron el lenguaje de programación R. Este lenguaje conforma un conjunto integrado de funciones que facilita la manipulación de datos, cálculo y visualización gráfica. En general, constituye un entorno en el cual se implementan técnicas estadísticas aplicables, ya que en el entorno integrado se pueden instalar más de ocho mil paquetes, creados por diferentes personas en el mundo entero (The R Foundation 2023), lo que facilita el uso de una infinida de posibilidades para realizar análisis de datos de múltiples temas.

Cierto que existen otros lenguajes de programación que permiten la aplicación de técnicas de minería de textos, como Python. Este lenguaje de programación es simple y fácil de aprender, incluye múltiples bibliotecas y herramientas en un solo lenguaje de programación (Python 2023). Según esto, R tiene múltiples ventajas para la visualización de datos, mientras que Python resulta sencillo y fácil de aprender. Esta revisión de literatura se enfocó en identificar artículos que usaron el lenguaje de programación R para aplicar la minería de textos, ya que esta investigación usó este lenguaje de progamación. Por ejemplo, Urbizagástegui-Alvarado (2022) utilizó la minería de textos en la construcción de encabezamientos de materia, palabras clave y/o términos de indexación para artículos de revistas con RStudio. Por igual, Contreras Barrera (2016) empleó la minería de textos para desarrollar un clasificador automatizado para la clasificación de material bibliográfico.

La mayoría de las investigaciones publicadas en bibliotecología y documentación que usan la minería de texto y el PLN no utilizan el lenguaje de programación R, sino otros softwares que permite realizar este tipo de análisis. En este punto toman fuerza y resultan trascendentes este modelo de artículos para la bibliotecología, debido a que muestran que R resulta de gran valor en dicha área del conocimiento, aunque siempre teniendo presente las limitaciones de los paquetes, los cuales se pueden cargar en este lenguaje y el conocimiento del lenguaje de programación, así como el uso de otros lenguajes de programación.

También se encontraron estudios que aplicaron la minería de textos, el PLN y el lenguaje de programación R en otras áreas del conocimiento que examinaron los temas explorados en el dominio de las ciencias sociales sobre investigaciones de COVID-19 (Roychowdhury, Bhanja y Biswas 2022); y los que desarrollaron un prototipo de visualización para apoyar el aprendizaje del análisis *netnográfico* con la exploración de colecciones de datos, utilizando métodos de análisis de red y minería de textos (Musabirov y Bulygin 2020); incluso, los que

analizaron la longitud y la frecuencia de la coocurrencia de palabras (n-gramas) en el contenido no estructurado de notas clínicas, mediante análisis proporcional y agrupamiento jerárquico no supervisado (Rahimian *et al*. 2019); además de los que estudiaron el léxico de palabras, con el fin de extraer opiniones de textos no estructurados publicados en Facebook y Twitter, con el propósito de identificar el discurso del odio (Udanor y Anyanwu 2019); y aquellos que compararon la citación con los tweets en la psicología (Ye y NA 2018).

En general, la aplicación de la minería de textos en la bibliotecología y la documentación constituye un tema novedoso que requiere de la exploración en los diferentes ámbitos de competencia de estas disciplinas. Esta revisión de la literatura mostró que la minería de textos ha sido aplicada en la organización de la información que se interpola con la recuperación de la información.

Metodología

Este artículo conforma un estudio exploratorio que utilizó los datos obtenidos en una búsqueda de información en Scopus en el campo de búsqueda: *Article title, Abstract, Keywords*, la cual se llevó a cabo el 22 de enero de 2023. En este campo se agregaron las palabras «Chocó AND Colombia»[1] que arrojó un resultado de 668 documentos. Se tomaron todas las referencias bibliográficas de los documentos indexados por Scopus, estos incluyen artículos científicos (628 documentos), capítulos de libro (11 referencias), ponencias (27 registros) y libros (dos ejemplares) que fueron publicados en los años que abarcan de 1913 a 2022.[2] El producto de la búsqueda fue exportado considerando la información bibliográfica de autor(es), título, publicación periódica, volumen, número y año, además se agregó el resumen. Los resultados se obtuvieron en un archivo de texto que incluyó la referencia bibliográfica y el resumen de cada artículo en inglés, este archivo se revisó y se eliminaron referencias, símbolos de *copyright* o resúmenes en español,

1 El operador booleano AND asocia dos términos (Colombia y Chocó), busca en este caso en los documentos que incluyan en título del artículo, resumen y palabras clave ambos términos. Por lo tanto, todos los documentos recuperados deberían incluir ambas palabras clave. Cabe aclara que las palabras clave pueden aparecer en el título, en el resumen, o bien, en las contenidas en los datos bibliográficos de los documentos indexados en Scopus.

2 El número de resultados arrojados por esta base de datos varía de acuerdo con la fecha de búsqueda; por ejemplo, si se realiza la búsqueda con los mismos parámetros usados que el 22 de enero de 2023, los resultados son 683 documentos, es decir, varían en cantidad. Cabe anotar que los documentos que se utilizaron en este artículo están contenidos en una base de datos en EndNote y al llevar a cabo las revisiones sugeridas en el proceso de evaluación los 668 documentos tratan sobre el Chocó, entiéndase que este lugar geográfico conforma un departamento o región geográfica. También, que el Chocó no tiene homónimos en otro lugar del planeta y que sólo está ubicado en Colombia. Cabe aclarar que se encontraron 13 documentos que únicamente incluyen la palabra clave «Chocó» y que tratan sobre este lugar geográfico de Colombia, es decir, no incluyen la palabra clave «Colombia», pero tratan sobre el tema de interés.

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2024.98.58833

también se cambiaron títulos que aparecieron en español por el título en inglés, ya que la mayoría de ellos aparecen indexados en este último idioma. Aunque es importante aclarar que los títulos de las revistas se dejaron en el idioma en el que aparecen. El archivo se convirtió en formato pdf y se manipuló con la ejecución de cuatro procesos que se llevaron con el software R de acceso libre en el entorno de desarrollo integrado (EDI). En el EDI se cargaron los paquetes o las librerías que contienen un conjunto de funciones, datos y códigos de R que permitieron manipular y convertir los datos en formatos legibles por los diferentes paquetes para obtener los resultados esperados en este trabajo. En cada proceso se explicó qué paquetería se usó, no se incluyó el código que se utilizó por la extensión que debe tener este artículo,[3] por eso se agregaron las referencias bibliográficas de cada uno de los paquetes para que cualquier otro investigador pueda replicar este estudio usando las funciones que establecen la paquetería para realizar minería de textos.

Con la aplicación del primer proceso se realizó la codificación del texto que consistió en preprocesar los documentos y almacenar la información en un archivo de texto sin formato. En esta etapa se usaron las librerías *pdftools* (Ooms 2023) que permitió extraer el texto del documento; *tm* (Feinerer y Hornik 2023) posibilitó realizar la minería de textos; *readr* (Wickham, Hester y Bryan 2023) se utilizó para leer archivos en diferentes formatos como pdf, txt, etc.; *Dplyr* (Wickham *et al.* 2023) se usó para analizar el subconjunto de datos o el *data frame*; y *tibble* (Müller y Wickham 2023) son *data frames* que ayudan a modificar algunas características antiguas de R Studio.

Con la aplicación del segundo proceso se obtuvo la lista de palabras y su frecuencia de aparición en el texto. Se aplicó la *tokenización* que consiste en dividir un documento en una secuencia de palabras y la eliminación de todos los signos de puntuación, marcas y reemplazo de las tabulaciones y otros caracteres que no son texto por espacios en blanco. La unión del conjunto de palabras y documentos de texto se denomina «diccionario de una colección de documentos». En esta etapa se usaron las paqueterías *tidyr* (Wickham y Romain 2016) que facilitó la creación de datos ordenados, es decir, cada columna representa una variable, cada fila una observación y cada celda tiene un valor único; *Purr* (Lionel y Wickham 2018) permitió realizar bucles o interacciones repetitivas y *stringr* (Wickham 2022) se usó para generar cadenas de texto.

Con la aplicación del tercer proceso se redujo el número de palabras contenidas en el diccionario de la colección de documentos, por ello se usaron métodos de filtrado, derivación y lematización. Con el filtrado se eliminaron las palabras vacías; por ejemplo, artículos, conjunciones, preposiciones, etcétera. Con

3 Los interesados en obtener el código usado en el lenguaje de programación R pueden solicitarlo al correo electrónico: crestrepoarango@gmail.com

la lematización se identificaron las formas verbales sin conjugar, sustantivos sin conjugar y palabras en singular. Con la derivación se construyeron las formas básicas de las palabras, es decir, se extrajo la raíz natural de un grupo de palabras, por eso se eliminan el plural, los sufijos en los sustantivos y los verbos como «ing» y otros sufijos. En esta etapa se usó *stopwords* (Muhr, Benoit y Watanabe 2023) para eliminar los artículos, los pronombres y otras palabras vacías.

Con la aplicación del cuarto proceso se obtuvo la lista de *bigrams* y su frecuencia de aparición en el texto, la coocurrencia de palabras y se aplicó el método de colocaciones. Este proceso consiste en utilizar software para manipular las palabras obtenidas e identificar patrones, pares de palabras, colocaciones y coocurrencia de palabras. En esta etapa se usaron las herramientas: *tidytext* (Silge 2023) para ordenar los datos obtenidos; *NLP* (Hornik 2022) para el procesamiento del lenguaje natural; *igraph* (Csardi y Nepusz 2006), *ggraph* (Pedersen 2022) y *scales* (Wickham y Seidel 2022) así como *ggplot2* (Wickham 2009), se emplearon para generar gráficos. También se usó *quanteda* (Benoit y Nulty 2016) para crear y administrar corpus textuales, extraer características de datos textuales y analizar esas características utilizando métodos cuantitativos (Mendoza 2016).

Resultados

Características generales de las palabras

Se examinó la frecuencia de aparición de las palabras en la información bibliográfica y resúmenes de los 668 documentos indexados en Scopus publicados entre 1913 y 2022. Se hallaron 997 palabras con una frecuencia de aparición de 1 268 a diez, estas palabras se agruparon en singular o plural, verbos en presente o pasado, adjetivos, conjunciones y sustantivos.

Con el proceso de filtrado, lematización y derivación se encontraron 219 términos que aparecen 1 268 a 49 veces, de este filtrado se identificaron 76 palabras que tienen entre 1 268 a 100 apariciones en el texto. Algunas de las palabras con el mayor número de frecuencia de aparición: «species», «forest», «pacific», «tropical», «diversity», «América», etcétera. (véase Anexo *Tabla 1*). Estas palabras representan las temáticas de investigación que se están desarrollando sobre el Chocó (Colombia).

Análisis de bigrams

Se identificaron los *bigrams* o pares de palabras. Se encontraron 89 841 pares de palabras que tienen una frecuencia de aparición de 221 a una vez. Destacan los

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2024.98.58833

bigrams que aparecen de 221 a 20 veces en el análisis; por ejemplo, «new species», «pacific coast», «pacific ocean», «rain forest», «San Juan», «gold mining», «Atrato river», «river basin», «Chocó biogeographic», y «climate change», etcétera. (véase Anexo *Tabla 2*). Estos pares de palabras evidencian cuáles son las temáticas de interés en la comunidad académica que está atraída en investigar acerca de este lugar geográfico.

Los *bigrams* fueron filtrados y se eliminaron las palabras vacías (artículos, pronombres y preposiciones), se obtuvieron 88 467 que fueron separados en dos columnas, y, a partir de esta separación, se hallaron 65 985 con una frecuencia de aparición de 133 a una vez. Algunas de las palabras que destacan: «biogeographic/climate/change, Pacific/Ocean, Biogeographic/Chocó», entre otras (véase Anexo *Tabla 3*).

Análisis de coocurrencias de palabras

La ocurrencia de palabras se entiende como la coexistencia de las palabras clave «a» y clave «b» que se usan juntas en el mismo documento. Su propósito radica en analizar las relaciones entre las palabras clave que se podrían considerar las más significativas por su frecuencia de aparición en un documento (Callon, Courtial y Laville 1991). A partir del filtrado de los pares de palabra, se extrajeron aquellas con una frecuencia de aparición en el texto mayor a diez. Se obtuvo un gráfico de co-ocurrencia de palabras que muestra las relaciones temáticas agrupadas por estas. Con dicha técnica se visualiza la estructura del conocimiento de un campo intelectual como es el caso de la literatura indexada en Scopus sobre el Chocó (Colombia) (véase *Figura 1*).

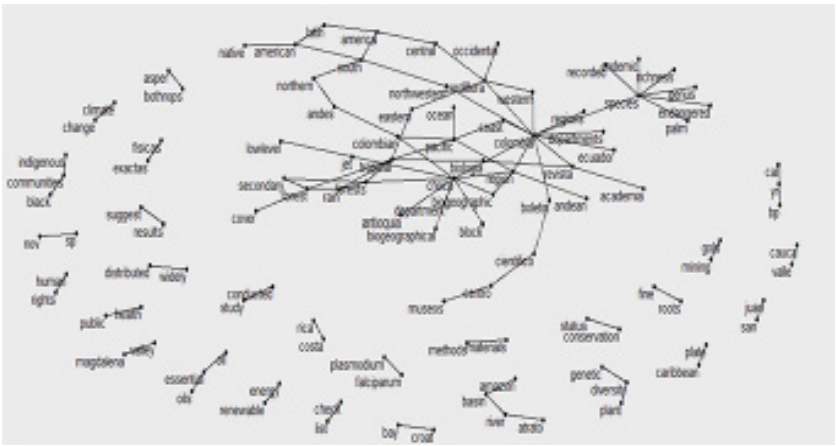

Figura 1. Coocurrencia de palabras sobre los artículos publicados del Chocó, Colombia
Fuente: Elaboración propia.

Con base en la coocurrencia de palabras se observan aquellas clave que están conectadas con otros subgrupos de palabras clave, se localizan en el centro de la *Figura 1*. También se encontraron temáticas periféricas, las cuales se localizan alrededor de dichas palabras que representan temáticas con mayor frecuencia de aparición, además no están interconectadas con otros subgrupos de palabras clave y normalmente se conectan dos o hasta seis palabras. Este análisis por igual muestra las temáticas que representan las preocupaciones de la comunidad científica interesada en investigar este lugar geográfico, incluso Urbizagastegui-Alvarado (2021) encontró en el análisis de la bibliometría brasilera las temáticas de interés en ese campo del conocimiento.

Análisis de colocaciones de palabras

Se analizaron las colocaciones de palabras que aparecen en el texto junto a «gold», «Chocó», «biodiversity» y «climate» para determinar las relaciones temáticas. Se estudiaron con el enfoque que introduce la noción de «núcleo», es decir, la palabra que es objeto de análisis y las palabras relacionadas que son aquellas que combinan con dicho núcleo. El principal requisito establece que estén relacionadas sintácticamente, es decir, cada palabra tiene una función dentro de una oración. Se considera que colocación no indica un fenómeno lineal, las colocaciones se pueden dar entre palabras que no están unidas. El núcleo puede estar separado por varias palabras a la derecha o a la izquierda (Corpas 2001).

La colocación de la palabra «gold» se relacionó de forma sintáctica con las palabras: «Choco», «mining», «mercury», «platinum», «alluvial», entre otras (véase *Figura 2*).

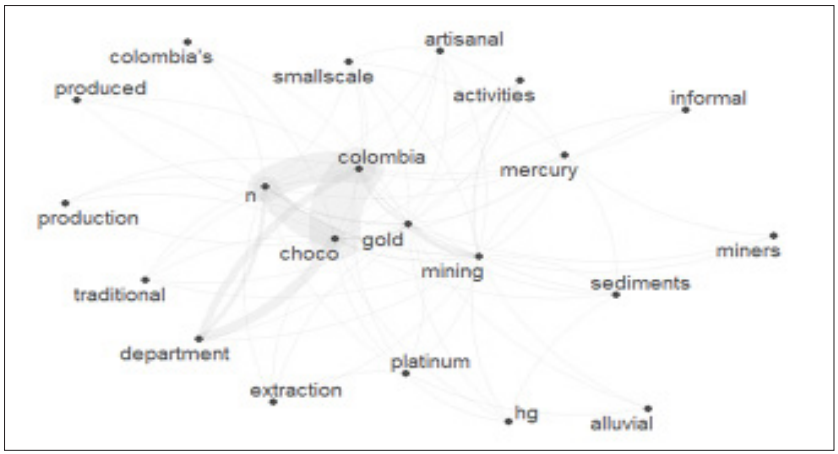

Figura 2. Colocación de la palabra «gold»
Fuente: Elaboración propia.

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2024.98.58833

La posición de la palabra «Chocó» se relacionó sintácticamente con las palabras: «Colombia», «pacific», «biogeographic», «coast», entre otras (véase *Figura 3*).

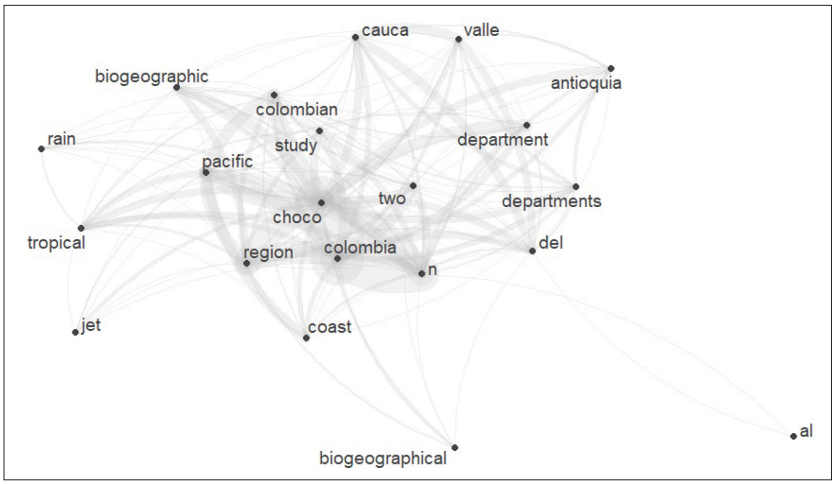

Figura 3. Colocación de la palabra "Chocó"
Fuente: Elaboración propia.

La disposición de la palabra «biodiversity» se relacionó de manera sintáctica con las palabras: «conservation», «tropical», «forest», «agricultural», «deforestation», «climate», etcétera (véase *Figura 4*).

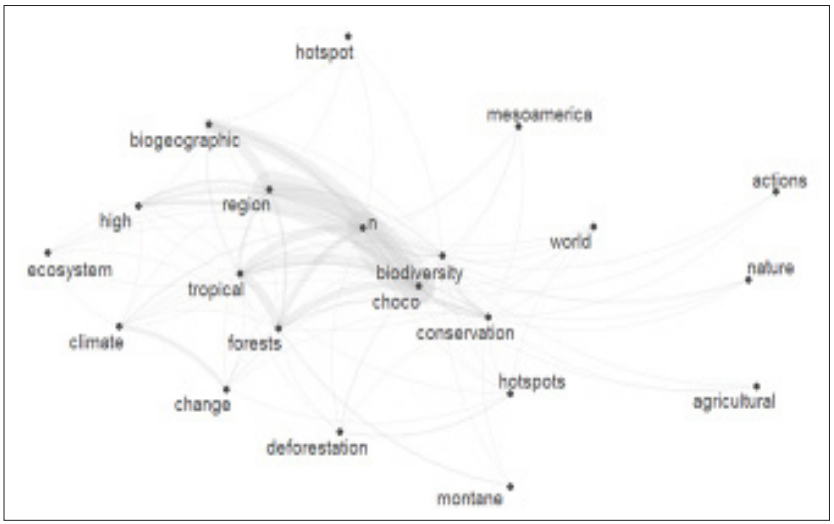

Figura 4. Colocación de la palabra «biodiversity»
Fuente: Elaboración propia.

En cambio, al colocar la palabra «climate» se relacionó con: «change», «forest», «basin», «sediment» y «marine», y demás (véase *Figura 5*).

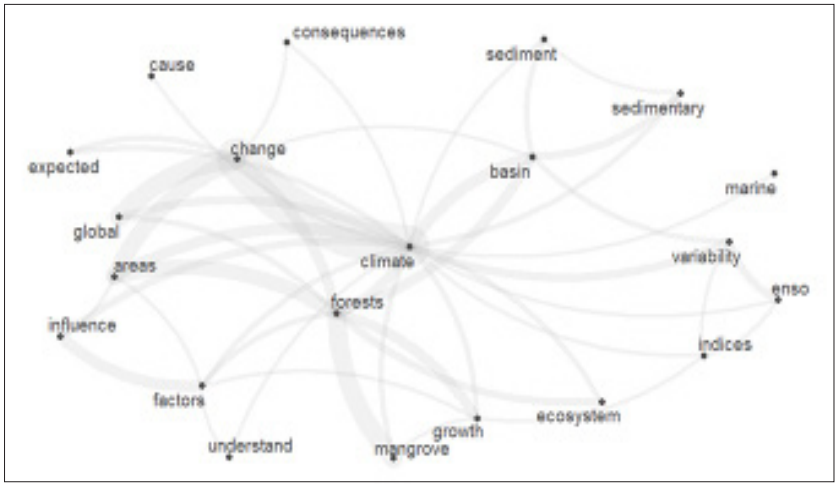

Figura 5. Colocación de la palabra «climate»
Fuente: Elaboración propia.

En general, la técnica de las colocaciones mostró las relaciones semánticas de las palabras núcleo. Las conexiones encontradas muestran no sólo un vínculo sintáctico, sino que presentan las preocupaciones de los investigadores que logran publicar sus trabajos en revistas indexadas en Scopus. También evidencian los intereses temáticos de esas revistas, es decir, en las indexadas en Scopus no se encontró un interés creciente por los artículos que tiene como objeto de estudio los problemas sociales que aquejan al Chocó en Colombia.

Consideraciones finales

La minería de textos indica una técnica para identificar tendencias temáticas a partir de la manipulación y conversión de un documento o conjunto de ellos en un archivo de texto, con el fin de aplicar métodos de filtrado, derivación y lematización que muestran las palabras más frecuentes en un documento o grupo de estos. Se encontró, en el caso de la literatura indexada en Scopus sobre Chocó (Colombia), las temáticas preponderantes de minería, biodiversidad, cambio climático, bosque tropical, océano pacífico, etcétera.

Las técnicas de minería de textos describen patrones que explican o resumen las relaciones subyacentes en los datos (Mariñelarena-Dondena, Errecalde y Cas-

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2024.98.58833

tro 2017). Se enfoca en el descubrimiento de «tendencias, desviaciones y asociaciones entre 'gran´ la cantidad de información textual» (Montes-y-Gómez 2000, 1). Con el apoyo de esta técnica se identificaron los pares de palabras o *bigrams*, entre los cuales destacan por el número de apariciones: «pacific coast», «biogeographic región», «pacific ocean», «rain forest», «San Juan», «gold mining», «tropical forest», «tropical pacific», «Atrato river», «river basin», «Chocó biogeographic», y «climate change», entre otras. Estas temáticas pueden convertirse en descriptores que representen el contenido de los artículos indexados en Scopus. La información extraída con la minería de texto y la PLN representan un importante insumo para la organización de la información en el campo de la bibliotecología y la documentación, pues las temáticas que se pueden usar para representar el contenido de un documento están inmersas dentro del mismo contenido analizado.

Para Eíto Brun y Senso (2004), entre las principales funciones de la minería de textos está identificar conceptos y crear redes de conceptos. Esto es, las palabras por sí solas no tiene un significado, deben estar rodeadas por otras para dar sentido a una oración, párrafo o texto. Principalmente el análisis de co-ocurrencia encuentra conceptos que aparecen juntos en un documento, o sea, existe una relación de proximidad de dos o más palabras en una frase, párrafo o documento. Por ejemplo, en el caso de «forest» y «rain»; «america» y «central»; «america» y «native», y así de manera sucesiva se encuentra una relación semántica entre los términos que están presentes en la *Figura 1*.

La base teórica de las colocaciones se fundamenta en que «conoces a una palabra por sus compañeras» (Firth 1957, 6), en resumen, el significado de una palabra se cimienta en el de las otras que la rodean. Esto quiere decir que están formadas por mínimo dos palabras (Corpas 2001). Estas muestran las relaciones de significado de palabras como «gold» con mercurio, por ejemplo; «Chocó» con biogeográfica; «biodiversity» con conservación; «climate» con cambio, y así sucesivamente. Estos constituyen algunos ejemplos de palabras que dan sentido a las que se utilizaron en este trabajo para las colocaciones.

En síntesis, la minería de textos es una técnica que se puede aplicar con el lenguaje de programación R, por medio de la paquetería y funciones incluidas en estos. Este lenguaje de programación permite extraer temáticas, o bien, patrones textuales para representar el contenido de un documento. Lo interesante de la minería de textos radica en que se puede aplicar no sólo en las redes sociales para extraer las palabras que representan las distintas emociones expresadas, por medio del lenguaje humano, sino en documentos académicos y técnicos que incluyen conceptos, lugares, acontecimientos, por ejejmplo, y que son usados por los científicos para explicar los fenómenos estudiados en las áreas del conocimiento. La extracción de estas palabras permite identificar tendencias temáticas y relaciones semánticas entre conceptos. Estas tendencias y relaciones muestran

el inicio, la marcha y el horizonte de las investigaciones, por medio del análisis de los datos bibliográficos y resumen de la literatura producida sobre el Chocó e indexada en Scopus.

Referencias

Al-Betar, M. A., Abasi, A. K., Al-Naymat, G., Arshad K. y Makhadmeh S. N. 2023. Optimization of scientific publications clustering with ensemble approach for topic extraction. *Scientometrics*, (128): 2819–2877. https://doi-org.biblioteca-colmex.idm.oclc.org/10.1007/s11192-023-04674-w.

Alkan, B. B., Karakuş L. y Direkci B. 2023. Knowledge discovery from the texts of Nobel Prize winners in literature: sentiment analysis and Latent Dirichlet Allocation. *Scientometrics*, (128): 5311–5334 (2023). https://doi-org.biblioteca-colmex.idm.oclc.org/10.1007/s11192-023-04783-6.

Benoit, K. y Nulty P. 2016. *quanteda*: Quantitative Analysis of Textual Data. Consultado 2 de agosto, 2023. https://CRAN.R-project.org/package=quanteda

Callon, M., Courtial J. P. y Laville F. 1991. Co-word analysis as a tool for describing the network of interactions between basic and technological research: the case of polymer chemistry. *Scientometrics*, 22: 155-205.

Csardi, G. y Nepusz, T. 2006. The igraph software package for complex network research. *InterJournal* Complex Systems, 1695. Consultado 2 de Agosto, 2023. https://igraph.org.

Contreras B., M. 2016. Minería de texto en la clasificación de material bibliográfico. *Biblios*, (64): 33-43. Consultado 4 de junio, 2023. https://www.redalyc.org/journal/161/16148511003/html

Contreras B., M. 2014. Minería de texto: una visión actual. *Biblioteca Universitaria*, 17 (2): 129-138.

Corpas P., G. 2001. En torno al concepto de colocación. *EUSKERA*, 46: 89-108.

Eíto B., R. y Senso, J. A. 2004. Minería textual. *El Profesional de la Información*, 13 (1): 11-27.

Firth, F. R. 1957. Modes of Meaning. *Papers in Linguistics,1934-1951*. London: Oxford University Press, p. 190-215.

Feinerer, I., K. Hornik. 2023. tm: Text Mining Package. R package version 0.711. Consultado 2 de agosto, 2023. https://CRAN.R-project.org/package=tm

Gobernación del Chocó. 2023. Información general. Quibdó: Gobernación. Consultado 2 de agosto, 2023. https://www.choco.gov.co/departamento/informacion-general.

Lionel, H. y Wickham H. 2018. Purrr: Functional Programming Tools. Consultado 2 de agosto, 2023. https://CRAN.R-project.org/package=purrr.

Hornik, K. 2022. Package nlp. Consultado 2 de agosto, 2023. https://cran.r-roject.org/web/packages/NLP/NLP.pdf

Hotho, A., A. Nürnberger y G. Paaß. 2055. A brief survey of text mining. *Journal for Language Technology and Computational Linguistics*, 20 (1): 19-62.

Hosseini, S., H. Baziyad, R. Norouzi, S. Jabbedari Khiabani, G. Gidófalvi, A. Albadvi, A. Alimohammadi y S. Seyedabrishami. 2021. Mapping the intellectual structure of GIS-T field (2008–2019): a dynamic co-word analysis. *Scientometrics*, (126): 2667-2688.

Mendoza V., J. B. 2016. Introducción a la minería de textos con R. RPubs. Consultado 2 de Agosto, 2023. https://rpubs.com/jboscomendoza/mineria-de-textos-con-r.

Ma, Yongchao, Ying Teng, Zhongzhun Deng, Li Liu y Yi Zhang Deng. 2023. Does writing style affect gender differences in the research performance of articles? An empirical study of BERT-based textual sentiment analysis. *Scientometrics*, (128): 2105–2143. https://doi-org.biblioteca-colmex.idm.oclc.org/10.1007/s11192-023-04666-w.

Mariñelarena-Dondena, L., M. L. Errecalde y A. Castro S. 2017. Extracción de conocimiento con técnicas de minería de textos aplicadas a la psicología. *Revista Argentina de Ciencias del Comportamiento*, 9 (2): 65-76.

Montes-y-Gómez, M. 2001. *Minería de texto: un nuevo reto computacional*. México: Instituto Politécnico Nacional. https://ccc.inaoep.mx/~mmontesg/publicaciones/2001/MineriaTexto-md01.pdf

Muhr, D., K. Benoit y K. Watanabe. 2023. stopwords: the R package. Consultado 2 de agosto, 2023. https://cran.r-project.org/web/packages/stopwords/readme/README.html

Müller, K. y H. Wickham. 2023. tibble: Simple Data Frames. Consultado 2 de agosto, 2023. https://tibble.tidyverse.org/.

Musabirov, I. y D. Bulygin. 2020. Prototyping text mining and network analysis tools to support netnographic student projects. International Journal of Emerging Technologies in Learning (iJET), 15 (10): 223-232.

Ooms, J. 2023. Package pdftools. Consultado 2 de agosto, 2023. https://cran.r-project.org/web/packages/pdftools/pdftools.pdf.

Pedersen, T. 2022. ggraph: An Implementation of Grammar of Graphics for Graphs and Networks. Consultado 2 de Agosto, 2023. https://github.com/thomasp85/ggraph

Python. 2023. El tutorial de Python. Consultado 18 de octubre, 2023. https://docs.python.org/es/3/tutorial/

Rahimian, M., J. L. Warner, S. K. Jain, R. B. Davis, J. A. Zerillo y R. M. Joyce. 2019. Significant and distinctive n-grams in oncology notes: a text-mining method to analyze the effect of OpenNotes on clinical documentation. *JCO Clinical Cancer Informatics*, (3): 1-9.

Roychowdhury, K., R. Bahanja y S. Biswas. 2022. Mapping the research landscape of Covid-19 from social sciences perspective: a bibliometric analysis. *Scientometrics*, 127 (8): 4547-4568.

Russell, M. A. 2013. *Mining the social web: data mining Facebook, Twitter, LinkedIn, Google+, GitHub, and more*. O'Reilly Media, Inc.

Shen, Si, Jiangfeng Liu, Litao Lin, Ying Huang, Lin Zhang, Chang Liu, Yutong Feng y Dongbo Wang. 2023. SsciBERT: a pre-trained language model for social science texts. *Scientometrics*, (128): 1241–1263. https://doi-org.biblioteca-colmex.idm.oclc.org/10.1007/s11192-022-04602-4.

Silge, J. 2023. Package tidytext. Consultado 2 de agosto, 2023. https://cran.r-project.org/web/packages/tidytext/tidytext.pdf.

The R Foundation. 2023. What is R? Consultado 18 de octubre, 2023. https://www.r-project.org/about.html

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2024.98.58833

Trask, A., D. Gilmore y M. Russell. 2015. Modeling order in neural word embeddings at scale. *Proceedings of the 32nd International Conference on Machine Learning*, 2266-2275. Lille, France: MLResearchPres.

Udanor, C. y Ch. C. Anyanwu. 2019. Combating the challenges of social media hate speech in a polarized society: a Twitter ego lexalytics approach. *Data Technologies and Applications*, 53 (4): 501-552.

Urbizagastegui-Alvarado, R. 2021. La bibliometría brasileña: minería de textos. *Revista ACB: Biblioteconomía em Santa Catarina*, 26 (1): 8-18.

Urbizagastegui-Alvarado, R. 2022. La minería de textos como subsidio para la organización de la información: un estudio exploratorio. *Revista Conhecimento em Ação*, 7 (2): 5-26.

Ye, Y. E. y J. C. Na. 2018. To get cited or get tweeted: a study of psychological academic articles. *Online Information Review,* 42 (7): 1065-1081.

Yin, X., H. Wang, P. Yin, H. Zhu y Z. Zhang. 2020. A co-occurrence-based approach of automatic keyword expansion using mass diffusion. *Scientometrics*, (124): 1885-1905.

Wickham, H., R. François, L. Henry, K. Müller y D. Vaughan. 2023. dplyr: a grammar of data manipulation. Consultado 2 de Agosto, 2023. https://github.com/tidyverse/dplyr.

Wickham, H., J. Hester y J. Bryan. 2023. readr: Read Rectangular Text Data. Consultado 2 de agosto, 2023. https://cran.r-project.org/web/packages/readr/index.html.

Wickham, H. 2022. Stringr: Simple, Consistent Wrappers for Common String Operations. Consultado 2 de agosto, 2023. https://cran.r-project.org/web/packages/stringr/index.html.

Wickham, H y D. Seidel. 2022. scales: Scale Functions for Visualization. Consultado 2 de Agosto, 2023. https://scales.r-lib.org. https://github.com/r-lib/scales.

Wickham, H. 2009. *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer-Verlag. https://link-springer-com.biblioteca-colmex.idm.oclc.org/book/10.1007/978-0-387-98141-3.

Wickham, Hadley y Francois Romain. 2016. dplyr: A Grammar of Data Manipulation. Consultado 2 de Agosto, 2023. https://CRAN.R-project.org/package=dplyr.

Zhang,Tingting, Baozhen Lee, Qinghua Zhu, Xi Han y Ke Chen. 2023. Document keyword extraction based on semantic hierarchical graph model. *Scientometrics*, (128): 2623–2647. https://doi-org.biblioteca-colmex.idm.oclc.org/10.1007/s11192-023-04677-7.

Para citar este texto:

Restrepo-Arango, Cristina. 2024. "Identificación de las temáticas de investigación del Chocó en la literatura indizada en Scopus". *Investigación Bibliotecológica: archivonomía, bibliotecología e información* 38 (98): 99-120.
http://dx.doi.org/10.22201/iibi.24488321xe.2024.98.58833

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2024.98.58833

Anexo 1

Términos	Frecuencia
species	1268
colombia	1131
choco	863
new	522
region	425
forest	332
colombian	306
two	304
pacific	301
study	284
results	230
tropical	217
data	215
diversity	207
america	204
areas	199
forests	198
one	198
distribution	186
found	185
western	178
south	176
three	176
high	172
using	170
also	169
analysis	167
genus	166
different	165
area	164

Términos	Frecuencia
malaria	145
regions	141
described	138
conservation	136
years	136
present	132
records	129
del	126
genetic	124
use	124
communities	117
river	117
biogeographic	116
ecuador	116
northern	116
among	114
caribbean	112
departments	110
total	110
within	110
groups	109
antioquia	108
number	108
cauca	107
known	107
showed	106
can	105
cases	105
environmental	105
well	105

population	162
used	160
first	158
populations	153
based	148
department	148
central	147
andes	146

eastern	104
local	104
american	103
cordillera	102
system	101
andean	100
associated	100
human	100

Tabla 1. Frecuencia de aparición de las palabras sobre Chocó, Colombia

No.	bigram	Frecuencia
1	new species	221
2	choco colombia	133
3	south america	107
4	sp nov	79
5	colombian pacific	69
6	pacific coast	66
7	m et	62
8	region colombia	59
9	choco region	58
10	two new	58
11	valle del	48
12	department choco	47
13	del cauca	45
14	revista de	45
15	western colombia	45

No.	bigram	Frecuencia
25	sp n	31
26	colombia ecuador	30
27	rain forest	29
28	san juan	28
29	northwestern south	27
30	climate change	26
31	south ame-rican	26
32	gold mining	25
33	species richness	25
34	choco department	24
35	zootaxa n	24
36	biogeographic choco	23
37	tropical forests	23
38	coast colombia	22
39	northern andes	22

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2024.98.58833

16	central america	39
17	biogeographic region	37
18	choco biogeographic	36
19	first time	36
20	new records	36
21	pacific region	35
22	described illustrated	34
23	antioquia choco	32
24	costa rica	31
40	species genus	22
41	tropical pacific	22
42	atrato river	22
43	regions colombia	22
44	river basin	21
45	biologia tropical	20
46	croat bay	20
47	endangered species	20
48	last years	20

Tabla 2. Bigrams sobre el departamento del Chocó, Colombia

Word uno	Word dos	Frecuencia
choco	colombia	133
south	America	107
sp	Nov	79
colombian	Pacific	69
pacific	Coast	66
region	colombia	59
choco	region	58
department	choco	47
western	colombia	45
central	america	39
biogeographic	region	37
choco	biogeographic	36
pacific	region	35
antioquia	choco	32
costa	rica	31
colombia	ecuador	30

rain	forest	29
san	juan	28
northwestern	south	28
climate	change	26
south	american	26
gold	mining	25
species	richness	25
choco	department	24
biogeographic	choco	23
tropical	forests	23
coast	colombia	22
northern	andes	22
species	genus	22
tropical	pacific	22
atrato	river	21
regions	colombia	21
river	basin	21
biologia	tropical	20
croat	bay	20
endangered	species	20

Tabla 3. Bigrams separados sobre el departamento del Chocó, Colombia