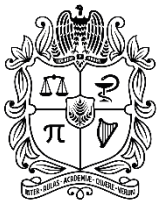


UNIVERSIDAD
NACIONAL
DE COLOMBIA

Análisis automático de contenido textual de colecciones de artículos científicos como apoyo al análisis documental y la gestión del conocimiento de una IES de la Orinoquía colombiana

Yerson Ferney Porras García

Universidad Nacional de Colombia
Facultad de Ingeniería, Departamento de Ingeniería de Sistemas e Industrial
Bogotá, Colombia
2021

Análisis automático de contenido textual de colecciones de artículos científicos como apoyo al análisis documental y la gestión del conocimiento de una IES de la Orinoquía colombiana

Yerson Ferney Porras García

Trabajo de grado presentado como requisito parcial para optar al título de:
Magister en Ingeniería - Ingeniería de Sistemas y Computación

Director:
Ph.D. Angel Alfonso Cruz Roa
Codirector:
Ph.D. Fabio Augusto González Osorio

Línea de Investigación:
Computación Aplicada y Ciencias de la Computación
Grupos de Investigación:
Machine learning, perception and discovery Lab (MindLAB) - Universidad Nacional de Colombia, Bogotá, Colombia
Grupo de Investigación en Tecnologías Abiertas (GITECX) - Universidad de los Llanos, Villavicencio, Colombia

Universidad Nacional de Colombia
Facultad de Ingeniería, Departamento de Ingeniería de Sistemas e Industrial
Bogotá, Colombia
2021

Dedicatoria

A mi madre por darme una vida llena de apoyo, fortaleza, resiliencia y aprendizaje.

A mi padre por las oportunidades de crecer que con la experiencia me otorgó.

A mis familiares y seres queridos que de alguna u otra forma fueron partícipes.

Por una vida llena de luz que me guíe hacia adelante y que nunca me permita dejar de soñar...

Yerson Porras

Agradecimientos

Gracias a Dios por todo. Agradecimientos a la Fundación CeiBA por la oportunidad brindada y el financiación de estudios de maestría, así como al convenio entre la Universidad de los Llanos y la Universidad Nacional de Colombia, Sede Bogotá para la oferta de la Maestría en Ingeniería de Sistemas y Computación en Villavicencio. Igualmente, gracias por el apoyo por parte del semillero Automatic Data-driven Laboratory de la Universidad de los Llanos, y a los grupos de investigación GITECX de la Universidad de los Llanos y MindLab de la Universidad Nacional de Colombia, Sede Bogotá.

Muchas gracias a mi director Ph.D. Angel Alfonso Cruz Roa y codirector Ph.D. Fabio Augusto González Osorio por todo el esfuerzo, tiempo, dedicación y dolores de cabeza que requerí para poder culminar este proyecto. Nuevamente, muchas gracias.

Gracias a mi madre Luz Mery García por el apoyo incondicional, a mi pareja Jessica López por ser esa fuente de fortaleza, paciencia y amabilidad que muchas veces necesité, a mis familiares y seres queridos que me aportaron algo y gracias a eso soy quien soy.

A todos, muchas gracias.

Resumen

La Universidad de los Llanos (Unillanos), con influencia en la Orinoquía colombiana, ha incrementado la producción de artículos científicos y realizar análisis del contenido textual de generación de nuevo conocimiento empieza a ser tedioso de forma manual. Este trabajo presenta un modelado y análisis automático de temas usando Latent Dirichlet Allocation (LDA) para el análisis semántico y temático de artículos científicos publicados por autores de la Unillanos disponibles en SCOPUS. LDA es comúnmente utilizado para descubrir relaciones de co-ocurrencia entre palabras y conformar grupos con valor semántico latente. Para el análisis, se obtuvo 137 artículos científicos en Inglés, analizados con LDA y usando la medida de desempeño Coherence Measure (CM). Se planteó un proceso sistemático de parámetros para determinar la parametrización del conjunto de datos y modelo LDA. El cual fue evaluado cuantitativa y cualitativamente. Se construyó un instrumento web para el diligenciamiento de las evaluaciones cualitativas por parte de un conjunto de expertos seleccionados, el cual se denominó "whatTopic". Cuantitativamente, el modelo obtuvo un valor de $CM = 0.639$ para un número de temas de 10. Cualitativamente, se comparó y relacionó temas propuestos con etiquetas preestablecidas por SCOPUS como "Scopus subrea" y "Scival Topic Prominence". En ambos casos se identificó como el principal tema la "Experimentación e Investigación" siendo transversal a los demás temas identificados que están asociados a características de la Unillanos como recursos naturales, agropecuarios, física y telecomunicaciones.

Palabras clave: Asignación Latente de Dirichlet; Modelado de Temas; Análisis de Texto; Gestión del Conocimiento; Artículos Científicos.

Abstract

The Universidad de los Llanos (Unillanos), with influence in the Colombian Orinoco region, has increased the production of scientific articles and the analysis of textual content from generation of new knowledge is difficult when done manually. This paper presents automatic topic modeling and analysis using Latent Dirichlet Allocation (LDA) for the semantic and topic analysis of scientific articles published by authors of Unillanos available in Scopus. LDA is commonly used to discover co-occurrence relationships among words and grouping with latent semantic value. For this analysis, 137 scientific articles in English were obtained, analyzed with LDA, and using the Coherence Measure (CM) as performance measure. A systematic parameter process was proposed to determine the parameterization of the dataset and LDA model. It was evaluated quantitatively and qualitatively. A web-based instrument was constructed for the completion of the qualitative evaluations by a group of selected experts, which was called "whatTopic". Quantitatively, the model obtained a value of $CM = 0.639$ for a number of topics of 10. Qualitatively, the proposed topics were compared and related to pre-established SCOPUS labels such as "Scopus subrea" and "Scival Topic Prominence". In both cases, "Experimentation and Research" was identified as the main topic, being transversal to the other identified topics associated with Unillanos characteristics such as natural resources, agriculture and livestock, physics and telecommunications.

Keywords: Latent Dirichlet Allocation; Topic Modeling; Text Analysis; Knowledge Management; Scientific Articles

Este Trabajo Final de maestría fue calificado en mayo de 2021 por el siguiente evaluador:

Juan David Suárez Moreno Msc.
Profesor Departamento de Ingeniería de Sistemas e Industrial
Facultad de Ingeniería
Universidad Nacional de Colombia

Contenido

Lista de Figuras

Lista de Tablas

1. Introducción

La generación de conocimiento es la relación existente entre conocimiento y las actividades racionales que lo producen [González Jiménez, 2001]. De igual forma, la apropiación de la generación de conocimiento, como forma de dinamizar la economía, influye en el desarrollo tecnológico y en la calidad de vida de la región de influencia [CONPES, 2018, Hernández-Sampieri et al., 2014]. Por lo tanto, la generación de nuevo conocimiento, como actividad intelectual y comercial, resulta de la Gestión del Conocimiento (GC) gracias al desarrollo de capacidades en la resolución de problemas a favor de la creación de ventajas competitivas [Osorio Núñez, 2003]. La GC en producciones científicas y académicas, generadoras de nuevo conocimiento, promueve la gestión en Investigación y Desarrollo (I+D), la evaluación competitiva del mercado, el impacto académico y científico, la gestión y adopción tecnológica y el crecimiento económico, entre otros [Feldman et al., 1998, Yoon and Park, 2004, COLCIENCIAS, 2015]. Por tal razón, países reconocidos como potencias económicas como por ejemplo Estados Unidos, Japón, Corea del Sur invierten en gran proporción, más del 2.7 %, recursos provenientes de su Producto Interno Bruto (PIB) en actividades de ciencia, tecnología e innovación (ACTI), Investigación y Desarrollo (I+D) y generación de conocimiento, a diferencia de Colombia con menos del 0.5 % [OCyT, 2020].

En 2017, Colombia presentó una tasa muy inferior de investigadores, considerando que por cada millón de habitantes, había 88 investigadores, a diferencia de México, Argentina y Dinamarca, con 315, 1,189 y 7,880 investigadores por cada millón de habitantes, respectivamente. Para América Latina y la Organización para la Cooperación y el Desarrollo Económicos (OCDE), los valores promedios fueron de 122 y 3,731 investigadores por cada millón de habitantes/respectivamente. En la figura **1-1** se evidencia que la cantidad de investigadores en Colombia es mínima en comparación con algunos países de referencia. Una posible causa, es la escasez de recursos e inversión con que cuentan los investigadores en sus respectivos países [CPC, 2019b].

El Índice de Complejidad Económica (ECI) indica cuantitativamente la intensidad relativa de conocimiento de una economía tomando en cuenta la intensidad de conocimiento de los productos que exporta [OEC, 2018]. Por lo cual, para el año 2018 en un ranking de 137 países, Colombia ocupaba el puesto 56 con un puntaje ECI de 0.23, mientras que Japón, Suiza y China Taipéi ocupaban los primeros lugares con un valor ECI de 2.17, 2.01 y 1.94, respectivamente. En cuanto a América Latina, lideraban México y Brasil ubicados en los

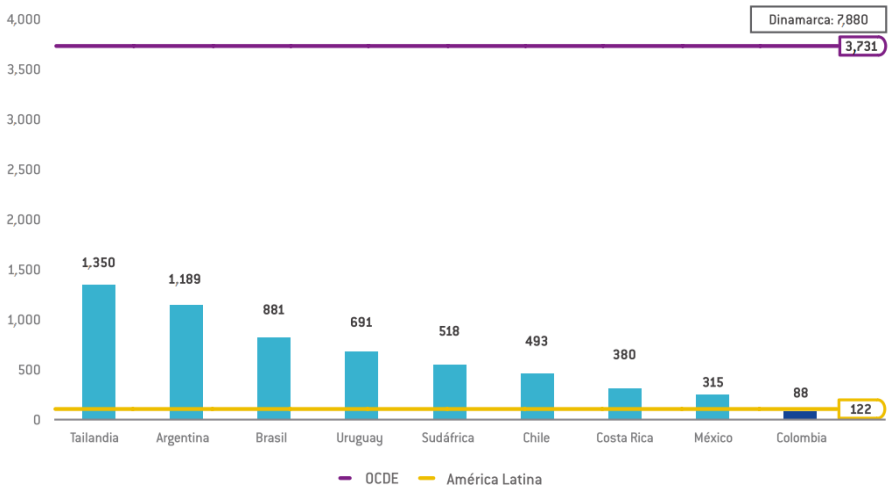


Figura 1-1.: Número de investigadores por cada millón de habitantes. Colombia y países de referencia, 2017. Tomado de [CPC, 2020].

puestos 21 y 39 con valores de 1.13 y 0.62, respectivamente. De igual forma, de los países líderes en número de investigadores por millón de habitantes, además de México, como son, Argentina y Dinamarca, su posición y valor ECI, corresponden a los puestos 57 y 23, con valores ECI de 0.22 y 1.03, respectivamente [OEC, 2018]. La figura **1-2** muestra el comportamiento de algunos países pertenecientes al ranking por medio de un histórico de valores ECI desde el año 2012 hasta el 2018. En dicha figura, es posible observar que los países desarrollados presentan un valor ECI promedio alto con comportamiento incremental, caso similar, en los países con mayor tasa de investigadores, a diferencia de los países en vía de desarrollo como Colombia, y en general América Latina, cuyo ECI es constante cercano a cero o desfavorable por su comportamiento descendente (valores ECI negativos) [OEC, 2018].

En Colombia se realizaron 114.495 publicaciones en revistas indexadas para el año 2019, un valor inferior a países vecinos como México, Argentina y Chile, con 247,369, 225,079 y 163,503 publicaciones respectivamente, siendo aún mayor la diferencia con Brasil que contó con más de un millón de publicaciones. A pesar de que Colombia ha logrado mejorar la calidad y aumentar el número de publicaciones permitiendo mejorar su posición al nivel mundial en el índice H[1]. Para el año 2019, Colombia invertía un 0.74 % y 0.28 % de su PIB en ACTI e I+D, respectivamente. Mientras que entre los países de América Latina con mayor inversión están Costa Rica con un 2.67 % para ACTI y Brasil con un 1.26 % para I+D [CPC, 2020, OCyT, 2020]. Por ejemplo, en la figura **1-3** se compara el porcentaje de inversión de Colombia con algunos países de la región, líderes en inversión a nivel mundial como Israel y Corea del Sur o Brasil a nivel regional, así como el valor promedio de la región. Es posible

[1]Según [CPC, 2020], "*El Índice H es el número de artículos de un país (h) que han recibido al menos h citaciones. Cuantifica la productividad científica de un país, así como su impacto.*"

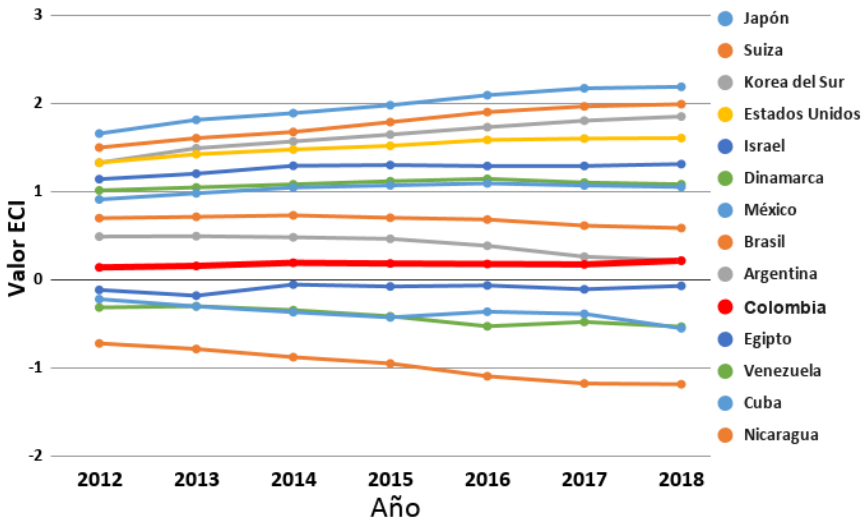


Figura 1-2.: Histórico de valores ECI de algunos países del ranking. Adaptado de [OEC, 2018].

observar una amplia diferencia en la inversión que realiza Colombia, y este valor muy inferior en comparación con el resto, siendo una de las posibles causas que generan déficit en ACTI, I+D y generación de nuevo conocimiento. Esto, tomando en cuenta que sin los suficientes recursos económicos que apoyen estos campos no se realizan investigaciones académicas y científicas que promuevan el desarrollo, la investigación, adopción y producción tecnológica y formación de investigadores como se ve en los países de referencia, razón por la cual, se evidencia en las necesidades y deficiencias de I+D del país.

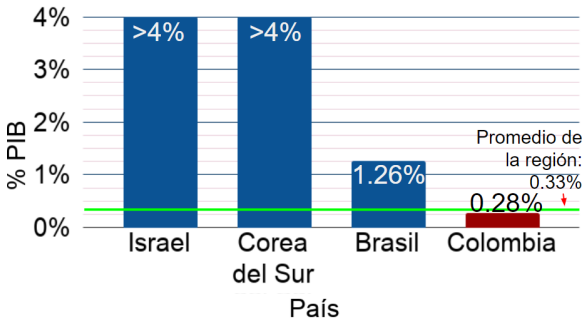


Figura 1-3.: Porcentaje de inversión de PIB para I+D. Colombia y países de referencia, 2019. Adaptado de [CPC, 2020, OCyT, 2020].

En el contexto nacional, la región de la Orinoquía y los Llanos Orientales, es una fuente importante de recursos naturales, biodiversidad y productos agropecuarios para Colombia [Estévez-Bretón, 2010], la cual ha venido desarrollando varios proyectos e investigaciones relacionadas [Soler-Tovar and Hernández-Rodríguez, 2018, CONPES, 2014]. Departamentos

como el Meta y Casanare, los cuales gracias a sus recursos naturales no renovables y producto de la explotación del petróleo y gas, cuentan con una importante cantidad de recursos económicos por regalías de dicha explotación en su presupuesto departamental [Departamento Nacional de Planeación, 2017]. Para el año 2019, el PIB de Colombia fue de casi $1.062 mil billones moneda legal [Banco Mundial, 2021], el presupuesto general de la nación fue de $258,997,305,209,927 moneda legal (aprox. 24.39 % del PIB) [Congreso de Colombia, 2018a]. De igual forma se decretó una suma de $18,564,591,529,959 moneda legal (aprox. 1.74 % del PIB) para el presupuesto del Sistema General de Regalías (SGR) para el bienio del 1 de enero de 2019 al 31 de diciembre de 2020 [Congreso de Colombia, 2018b]. Del SGR se asignó un total de $2,108,104,050,130 (aprox. 0.198 % del PIB) para el Fondo de Ciencia, Tecnología e Innovación (FCTeI) nacional. En cuanto a los departamentos de Arauca, Casanare y Meta, el SGR asignó las sumas de $413,709,058,167, $813,184,290,580 y $1,394,333,580,767 moneda legal, respectivamente. De dichas sumas se asignaron los siguientes montos para el FCTeI de cada departamento, para Arauca $36,711,437,126, Casanare $40,175,981,111 y Meta $42,477,883,300 moneda legal [Congreso de Colombia, 2018b].

Sin embargo, Arauca, Casanare y Meta han mostrado atraso en ciencia, tecnología e innovación en comparación con Bogotá D.C y Cundinamarca, Antioquia y Santander según el Índice Departamental Competitividad (IDC) y el Índice Departamental de Innovación para Colombia (IDIC) para el año 2019. El IDC permite evidenciar la calidad en el diseño e implementación de políticas públicas y de iniciativas privadas para el mejoramiento del bienestar de los ciudadanos, y el IDIC captura la complejidad y la multiplicidad de factores que intervienen en la innovación de la región por medio de enfoques que permiten identificar retos que cada departamento enfrenta en pro del desarrollo económico y social basado en la productividad, la competitividad y el crecimiento con equidad social [CPC, 2019a, DNP, 2020]. Esto es relevante, tomando en cuenta que el desarrollo de producción científica y generación de nuevo conocimiento, por medio de la formación de talento humano de alto nivel (maestría y doctorado), producción de artículos científicos y demás productos resultados de investigación reconocidos por el Sistema Nacional de Ciencia y Tecnología (SNCyT) del Ministerio de Ciencia, Tecnología e Innovación (MinCiencias) generados en las Instituciones de Educación Superior (IES) y demás centros de investigación, promueven el desarrollo tecnológico e innovación aumentando el IDC y el desarrollo económico de la región, incluso del país, puesto que los países más competitivos presentan mejores ingresos, mayor igualdad de oportunidades y sus habitantes están más satisfechos con la calidad de vida [COLCIENCIAS, 2015, COLCIENCIAS, 2018a, Ministerio de Educación de Colombia - Mineducación, 2010, COLCIENCIAS, 2018b, CPC, 2019b].

Es importante para una institución identificar y comprender su producción científica y la generación de nuevo conocimiento, ya sea para la toma de decisiones, dinamización de áreas y campos rezagados, fortalecimiento de áreas de mayor impacto y alcance, entre otros. Sin

embargo, a medida que se realizan investigaciones y se promueve la producción científica, se genera una gran cantidad de documentos con una amplia heterogeneidad temática que dificulta la identificación y el análisis documental. Las instituciones pueden realizar procesos de análisis documental de forma manual requiriendo inversión de recursos adicionales constantemente, como alternativa, estos procesos podrían automatizarse y así identificar los campos de acción y las áreas temáticas de su producción, con mayor eficiencia [Yau et al., 2014, Arrivillaga et al., 2016, Kong et al., 2017].

En Colombia, el 95.6 % de los investigadores están vinculados a IES y para aumentar la calidad e impacto de las inversiones y de la generación de nuevo conocimiento, es de vital importancia gestionar estímulos a la productividad académico-investigativa y promover el fortalecimiento de la institucionalidad por medio de generación y la protección de la propiedad intelectual y la propiedad industrial [CPC, 2019b]. La producción académica-científica de las IES como tesis, informes finales de trabajo de grado y, principalmente, artículos científicos es el resultado de investigaciones y fuente principal de generación de nuevo conocimiento de alto nivel, siendo las bases en el desarrollo tecnológico y académico para la innovación y mejora en el ámbito económico y productivo [CPC, 2019a]. Esta producción científica implica procesos de investigación que son sintetizados en documentos cuya cantidad de unidades va en continuo crecimiento generando, en la mayoría de los casos, variabilidad temática según los procesos y enfoques de investigación que tenga cada IES. Como enfoque para la aplicación y gestión del conocimiento, las IES requieren identificar sus campos de acción e investigación y así conocer su impacto en la sociedad, por lo cual, han realizado procesos (manuales o automáticos) de análisis textual sobre su producción científica (cienciometría o bibliometría) [Yau et al., 2014, Arrivillaga et al., 2016, Kong et al., 2017, CPC, 2019a]. Para el proceso manual, se requiere inversión y mano de obra de personas idóneas que sean capaces de realizar dicho proceso enfocados en la recuperación y extracción de información [Hamilton Wilson and Pezo Paredes, 2005]. Por otra parte, existen técnicas y herramientas computacionales que permiten automatizar el proceso de análisis textual requiriendo menos inversión de tiempo y personas especializadas en proporción al aumento del número de documentos. Dichos procesos computaciones permiten a una máquina (computador) identificar y "aprender" automáticamente relaciones e información de patrones de interés oculta a primera vista, este tipo de aprendizaje se conoce como aprendizaje automático o aprendizaje de máquina [Alpaydin, 2014].

En el área de las ciencias de la computación, el aprendizaje automático (*machine learning*) se define como un conjunto de modelos matemáticos cuyo objetivo radica en obtener información de forma automática basados en análisis estadísticos. Dichos modelos, tienen la capacidad de realizar inferencias basados en muestras y datos procesados permitiendo identificar, detectar, clasificar y predecir sobre otros nuevos conjuntos de datos [Dutton and Conroy, 1997, Alpaydin, 2014]. Uno de los campos de acción del aprendizaje automático es

la Minería de Texto (MT), que consiste en implementar técnicas automáticas para analizar y procesar datos de tipo textual, ya sean estructurados (e. g. bases de datos) o no estructurados (e. g. documentos, artículos científicos, blogs web, emails, entre otros) [Bhardwaj and Khosla, 2017].

Este estudio, con base en trabajos previos, plantea que la GC se puede desarrollar por medio de MT en documentos académicos como son los artículos científicos gracias a que la MT permite implementar técnicas computacionales enfocadas en reconocer patrones característicos de datos de tipo textual [Allahyari et al., 2017, Bhardwaj and Khosla, 2017, Pinto-Prieto et al., 2012, Sathya and Rajendran, 2013]. Adicionalmente, que la implementación de herramientas computacionales para la generación de nuevo conocimiento y GC promueven la adopción tecnológica y gestión de recursos tanto intelectuales (académico-investigativos) como tecnológicos [CPC, 2019b, CONPES, 2014, CONPES, 2018, CPC, 2019a]. Por último, que el uso de dichas herramientas enfocadas en la producción e impacto de la generación de nuevo conocimiento en IES, como la Universidad de los Llanos, institución de educación en investigación líder en la región de la Orinoquía, repercute en el desarrollo e innovación influyentes en el sector productivo de la región aumentando así, el desarrollo económico, competitividad y el apoyo tecnológico para la toma de decisiones [U-Sapiens, 2020, Andreu and Sieber, 1999, Osorio Núñez, 2003, Yoon and Park, 2004].

1.1. Definición del problema

A pesar de la amplia importancia de la GC y del auge tecnológico de herramientas de apoyo, en muchos casos el etiquetado de metadatos en los documentos académicos como trabajos de grado, tesis y artículos científicos, previos a algún tipo de análisis, se realiza de forma manual, significando una inversión en tiempo y mano de obra especializada considerables [Hamilton Wilson and Pezo Paredes, 2005]. Por su parte, las IES, como la Universidad de los Llanos, llegan a ser fuentes generadoras de documentos académicos y científicos que terminan siendo almacenados sin análisis y seguimiento por lo que se pierde todo el potencial de ser fuente de información y conocimiento en, por ejemplo, la toma de decisiones. Esto no permite generar adecuadamente una apropiación social y transferencia tecnológica del ámbito académico al sector productivo y económico [Feldman et al., 1998, Yoon and Park, 2004]. Por lo tanto, la MT se considera una técnica que puede apalancar la GC [Bhardwaj and Khosla, 2017, Pinto-Prieto et al., 2012, Sathya and Rajendran, 2013].

Finalmente, este trabajo de grado se abordó con el propósito de responder la pregunta de investigación principal ¿Cómo realizar análisis automático basado en contenido textual a partir de colecciones de artículos científicos como apoyo a la gestión de conocimiento en una IES de la región de la Orinoquía como lo es la Universidad de los Llanos?

1.2. Objetivos

1.2.1. Objetivo General

Implementar un método de aprendizaje computacional del estado del arte para el análisis automático de contenidos textuales, desde un enfoque de *topic modeling*, en artículos científicos como apoyo al análisis documental y gestión de conocimiento en una IES de la región de la Orinoquía.

1.2.2. Objetivos Específicos

- Consolidar el corpus de artículos científicos relevantes de una IES de la Orinoquía a través de fuentes de datos disponibles para análisis de contenido textual con *topic modeling*.
- Implementar un método computacional para el procesamiento y análisis de artículos científicos basado en un método de aprendizaje computacional de *topic modeling*.
- Evaluar cuantitativamente el desempeño computacional del método propuesto de aprendizaje computacional para el análisis de texto.
- Analizar los resultados obtenidos por el método propuesto.

1.3. Contribuciones y productos académicos

Entre las contribuciones del trabajo se encuentran:

- Modelo de temas entrenado disponible y su configuración de parámetros. El proceso de entrenamiento del modelo se describe en el Capítulo 4: Modelado y análisis de temas de artículos científicos.
- Conjunto de datos de artículos científicos de la Universidad de los Llanos en cada una de sus etapas de procesamiento. El conjunto de datos se discute en el Capítulo 3: Preparación y construcción del conjunto de datos.
- Software de acceso web que sirve como instrumento para la validación por expertos de un modelo de temas entrenado, denominado "whatTopic v1.4". Código fuente alojado en un repositorio web[2]. Capturas del instrumento de pueden observar en el anexo F. El diseño del instrumento se discute en la subsección 4.4.2 del Capítulo 4: Modelado y análisis de temas de artículos científicos.

[2]Repositorio: `https://bitbucket.org/Yerfer/tesisyporras/src`

- Reporte y análisis cuantitativo y cualitativo de los temas predominantes en la producción e investigación de la Universidad de los Llanos generadores de nuevo conocimiento. El reporte se puede ver anexo C. El análisis se discute en la Capítulo 6: Discusión y análisis.

1.4. Organización del documento

El documento se encuentra organizado de la siguiente manera, en el Capítulo 2: Marco conceptual y trabajos previos, se define formal y conceptualmente el área de conocimiento e investigación, problema, enfoque, método computacional y medida de desempeño de la implementación realizada en este trabajo, así como una presentación y síntesis de trabajos previos relacionados y el estado del arte en el área de investigación de modelado de temas. En el Capítulo 3: Preparación y construcción del conjunto de datos, se detalla el proceso metodológico de búsqueda y selección de artículos científicos, el proceso de preparación de los datos, la construcción de la representación textual implementada y la descripción del conjunto de datos consolidado. En el Capítulo 4: Modelado y análisis de temas de artículos científicos, se describe la metodología y búsqueda sistemática de parámetros óptimos del método de modelado de temas implementado a partir del método de aprendizaje computacional seleccionado, Latent Dirichlet Allocation (LDA), el proceso de exploración y visualización de temas latentes. En el Capítulo 5: Evaluación y resultados, se presentan y evalúan los resultados obtenidos por la evaluación cuantitativa y cualitativa. En el Capítulo 6: Discusión y análisis se interpretan, comparan y discuten los resultados obtenidos de la implementación del modelo de temas, tanto cuantitativos como cualitativos. En el capítulo 7: Conclusiones y trabajo futuro, se presentan las conclusiones, recomendaciones y trabajo futuro correspondientes. Por último, se relacionan las referencias bibliográficas empleadas en la construcción de este trabajo.

2. Marco conceptual y trabajos previos

Este capítulo presenta un marco de conceptos relacionados basado en la revisión de literatura realizada. Iniciando con la Minería de texto y su definición general, seguido del enfoque abordado de Modelado de temas y sus métodos tradicionales y de evolución de temas; y por último, la métrica de desempeño usada en este trabajo siendo la Medida de Coherencia. Seguidamente, se presenta un análisis y síntesis de los trabajos relacionados que abordaron el modelado de temas enfocados en artículos científicos.

2.1. Minería de Texto (MT)

El procedimiento en el que los algoritmos de aprendizaje automático son aplicados a datos de tipo textual, ya sean estructurados (e.g. bases de datos) o no estructurados (e.g. blogs web, documentos académicos o científicos, *emails*), se denomina Minería de texto (MT) [Bhardwaj and Khosla, 2017]. La MT busca encontrar y extraer patrones relevantes de forma automática ya que las computadoras no pueden "entender" textos no estructurados como lo puede hacer el ser humano [Allahyari et al., 2017, Sathya and Rajendran, 2013]. Algunas de las tareas de MT son: la recuperación de información, la cual busca encontrar documentos relevantes de una base de datos a partir de consultas de usuarios; la extracción de información, que se enfoca en reconocer y extraer características relevantes a partir de documentos textuales; la realización de resúmenes de textos, por la cual se genera un resumen automáticamente sintetizando el tema principal de un documento o un conjunto de estos; y el filtrado de información, que consiste en mostrar al usuario información específica y relevante para él [Allahyari et al., 2017, Manning et al., 2008, Hanani et al., 2001].

Por otra parte, el procesamiento del lenguaje natural (*Natural Language Processing* - NLP) consiste en herramientas, *software* o métodos de MT dotados de características funcionales que permiten el procesamiento del lenguaje natural humano, cuyo fin es extraer resultados, como características y patrones, basados en la información procesada proveniente de documentos escritos en lenguaje natural [Kurdi, 2016, Voorhees, 1999]. Algunas técnicas van desde cuantificar frecuencias de palabras para representar estilos y similitudes de documentos (e. g. Bag of Words – BOW), hasta tareas de "entendimiento" de expresiones humanas (e. g. análisis de sentimientos, minería de opinión, modelado de temas, entre otros). El NLP puede ser clasificado principalmente como una tarea de extracción de información, aunque, depende en gran medida del enfoque y del problema abordado [Bird et al., 2009, Allahyari

et al., 2017]. Para el desarrollo de este trabajo, se adoptó la extracción de información como enfoque de NLP.

2.2. Modelado de temas (Topic modeling - TM)

Modelado de temas (*Topic modeling - TM*) es un problema de MT enfocado al procesamiento del lenguaje natural (Natural Language Processing - NLP) que busca encontrar grupos representativos de palabras predominantes de un conjunto de documentos que permitan inferir temas latentes de los mismos documentos, permitiendo así conocer el área, investigación o tema abordado por cada uno [Allahyari et al., 2017, De Battisti et al., 2015, Bhardwaj and Khosla, 2017]. El modelado de temas es aplicable como medio para la GC permitiendo identificar áreas de investigación pertinentes en un conjunto de datos conformado por documentos académicos, como lo son los artículos científicos. Esto es posible gracias a que la MT permite implementar técnicas computacionales enfocadas en reconocer patrones característicos de datos de tipo textual [Allahyari et al., 2017, Bhardwaj and Khosla, 2017, Pinto-Prieto et al., 2012, Sathya and Rajendran, 2013, Aigneren, 1999, Feldman et al., 1998, García Clausó, 1994]. Típicamente se utiliza dos enfoques distintos: la factorización matricial basados en frecuencias de términos y operaciones entre matrices y los modelos probabilísticos basados en distribuciones de probabilidad a partir de supuestos bayesianos. El TM permite determinar tipos de agrupaciones "suaves"donde cada documento posee una distribución de probabilidad en todos los grupos (temas), caso contrario a la agrupación "estricta"de documentos. En el modelado de temas cada tema puede representarse como una distribución de probabilidad de palabras y cada documento se expresa como una distribución de probabilidad de temas. Por tanto, un tema es similar a un grupo de palabras y la pertenencia de un documento a un tema es probabilística [Allahyari et al., 2017, Alghamdi and Alfalqi, 2015].

Existen varias técnicas para abordar el problema de modelado de temas, algunos de los más relevantes y utilizados para el análisis de texto son: Probabilistic Latent Semantic Indexing (PLSI), Probabilistic Latent Semantic Analysis (PLSA), Latent Semantic Analysis (LSA), Latent Dirichlet Allocation (LDA), Correlated Topic Models (CTM) y Hierarchical Latent Dirichlet Allocation (Hierarchical LDA), Hierarchical Dirichlet Process (HDP), los cuales pueden variar tanto en los supuestos y formulación matemática, enfoques o áreas de aplicación [Hofmann, 1999, Hofmann, 2001, Blei et al., 2003, Blei and Lafferty, 2004, Yau et al., 2014]. LSA y el PLSA son modelos basados en frecuencias, en donde PLSA, calcula probabilidades basadas en frecuencias detectadas o parámetros disponibles en un conjunto de datos. En cambio LDA, CTM, Hierarchical LDA y HDP son modelos probabilísticos. En el caso de LDA, se enfoca en optimizar el cálculo de la probabilidad posterior (*maximum a posteriori* - MAP) cuyas frecuencias pueden llegar a ser desconocidas en un conjunto de datos futuro, por lo cual, opta por el enfoque bayesiano orientado al cálculo de la incertidumbre [Barreto, 2019]. En la figura **2-1** es posible observar una representación gráfica del proceso de mode-

lado de temas, de forma general, donde a partir de una colección de documentos se detectan y asignan grupos de términos denominados temas.

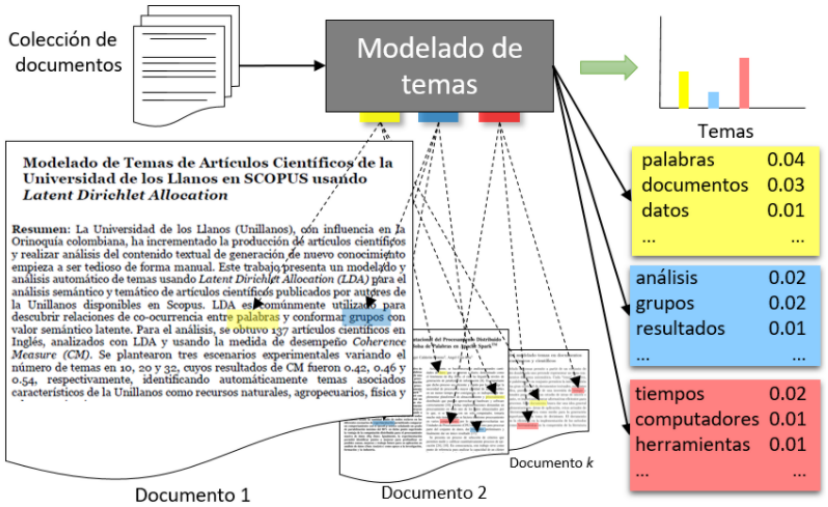


Figura 2-1.: Representación general de modelado de temas. Adaptado de [Blei, 2012].

De igual forma es posible clasificar los métodos de TM en dos grandes grupos según el alcance de la tarea que realizan: métodos tradicionales y métodos de evolución de temas [Alghamdi and Alfalqi, 2015]. Para el desarrollo de este trabajo se adoptó el modelado de temas por medio de modelos probabilísticos tradicionales.

2.2.1. Métodos tradicionales

Permiten detectar los temas de interés y áreas de investigación, sin embargo, no tienen la capacidad de evaluar la evolución, dependencia y relación temporal entre temas [Alghamdi and Alfalqi, 2015]. Existen métodos tradicionales basados en expresiones matriciales (e. g. LSA) y en modelos generativos probabilísticos (e. g. PLSA, LDA y CTM). Estos métodos han sido ampliamente usados en el análisis de artículos científicos con aplicaciones como detección de temas latentes y relaciones entre temas. De los anteriores métodos, LDA surgió con el objetivo de mejorar la forma de los modelos mixtos que capturan la intercambiabilidad tanto de las palabras como de los documentos en PLSA y LSA [Alghamdi and Alfalqi, 2015]. De hecho LDA hoy en día es el método tradicional más implementado y eficiente en modelado de temas en artículos académicos y científicos [Alghamdi and Alfalqi, 2015, Boyd-Graber et al., 2017, Chen et al., 2019].

2.2.2. Métodos de modelos de evolución de temas

Los modelos de evolución de temas (o *Topic Evolution Model* en Inglés) son aquellos métodos que modelan temas considerando un factor tiempo importante, ya sea tiempo discreto (e. g.

Dynamic Topic Models - DTM, *Multiscale Topic Tomography Model - MTTM* y *Dynamic Topic Correlation Detection - DCTM*), tiempo continuo (e. g. *Topic Over Time - TOT*), así como interrelaciones de citas o combinaciones entre sí. Estos modelos son usados principalmente con el objetivo de detectar la evolución de los temas en literatura científica y descubrir topologías de temas. Algunos métodos como TOT son derivaciones de LDA [Alghamdi and Alfalqi, 2015]. Los métodos tradicionales han sido usados en artículos científicos cuyas implementaciones no requieren considerar el factor tiempo. Permitiendo generar modelos de temas base viables para estudios posteriores. Por otro lado, los métodos de evolución de temas, debido al análisis temporal y procesos evolutivos y ontológicos, son más precisos en el descubrimiento de temas en comparación con los métodos tradicionales, aunque requieren una mayor cantidad de datos por cada unidad de tiempo para tener resultados relevantes [Alghamdi and Alfalqi, 2015]. Una representación resumen de la revisión de literatura se muestra en la figura **2-2**. En este trabajo, de acuerdo a esta revisión de literatura y el conjunto de datos objetivo se optó por un método tradicional de modelado de temas usando el más implementado y eficiente para artículos académicos y científicos, el cual es LDA.

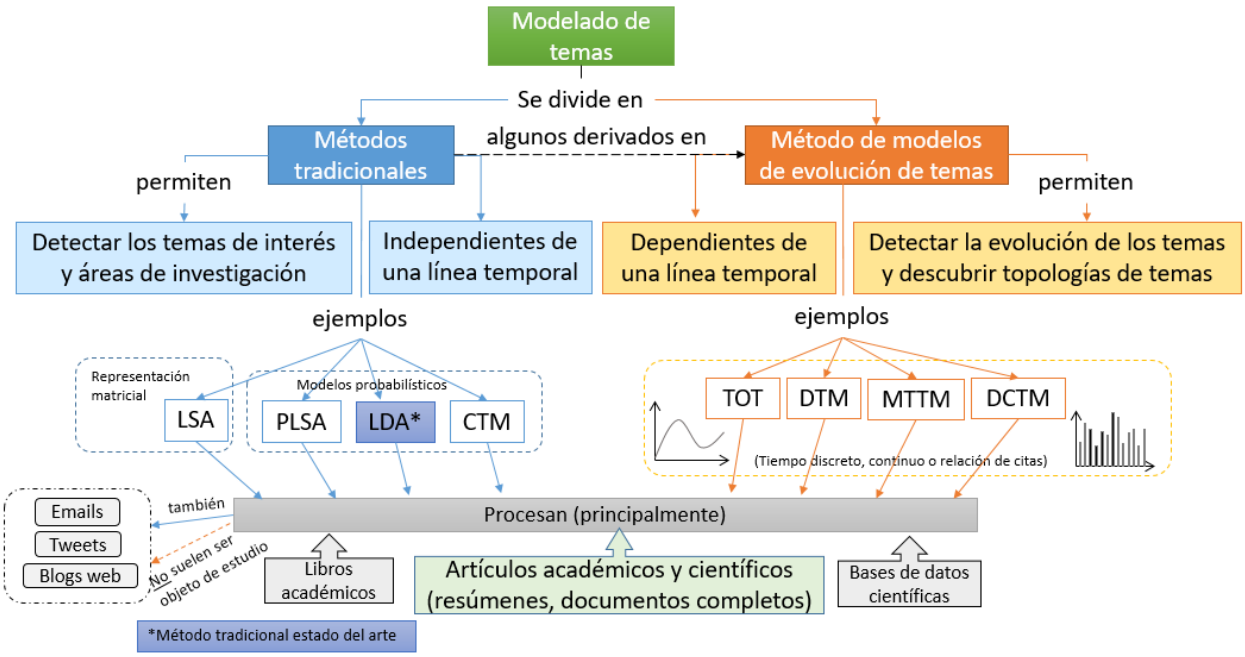


Figura 2-2.: Síntesis revisión de literatura modelado de temas en artículos académicos y científicos. Elaboración propia.

2.3. Medida de Coherencia (Coherence Measure - CM)

Según [Syed and Spruit, 2017], "*la coherencia de un tema se usa como un indicador de calidad del tema, basado en la hipótesis de distribución el cual señala que las palabras con significado similar tienden a co-ocurrir en un contexto similar*". La medida de coherencia (Coherence Measure – CM) es comúnmente usada para verificar la calidad de temas latentes obtenidos por algún método de modelado de temas, ya que permite cuantificar el grado de interpretabilidad humana de los temas [Syed and Spruit, 2018]. Entre las variantes de las medidas de coherencia estudiadas en [Röder et al., 2015], la usada en este trabajo es la definición de [Syed and Spruit, 2018], denominada como Cv, la cual representa una mayor relación de proporcionalidad directa entre valor cuantificable e interpretabilidad humana y puede tomar valores entre 0.0 y 1.0. La medida Cv recupera los recuentos de coocurrencia para las palabras dadas utilizando una ventana deslizante con un tamaño de ventana determinado. Los recuentos se utilizan para calcular el *normalized pointwise mutual information* (NPMI) de cada palabra principal con respecto a todas las demás, véase la ecuación 2-1, lo que da lugar a un conjunto de vectores de contexto $\overrightarrow{v}$, uno por cada palabra principal, véase la ecuación 2-2. La segmentación de un conjunto de palabras principales conduce al cálculo de la similitud entre cada vector de palabras principales y la suma de todos los vectores de palabras principales. Para esto se utiliza la medida de similitud coseno, véase la ecuación 2-3. El valor de coherencia Cv es la media aritmética de estas similitudes. La medida de coherencia (global o general) del modelado de temas es la media aritmética de los valores individuales de coherencia por tema [Röder et al., 2015].

$$NPMI(w_i, w_j)^{\gamma} = \left(\frac{\log \frac{P(w_i,w_j)+\epsilon}{P(w_i)\cdot P(w_j)}}{-\log(P(w_i, w_j) + \epsilon)} \right)^{\gamma} \tag{2-1}$$

$$\overrightarrow{v}(W') = \left\{ \sum_{w_i \in W'} NPMI(w_i, w_j)^{\gamma} \right\}_{j=1,\ldots,|W|} \tag{2-2}$$

$$\phi_{S_i}(\overrightarrow{u}, \overrightarrow{w}) = \frac{\sum_{i=1}^{|W|} u_i \cdot w_i}{\|\overrightarrow{u}\|_2 \cdot \|\overrightarrow{w}\|_2} \tag{2-3}$$

La relación entre palabras w_i y w_j de un vector de contexto en W se calcula con NPMI, véase ecuación 2-1, usando una constante ϵ para prevenir logaritmos de cero y γ para para dar más peso a los valores más altos de NPMI. Partiendo que, $\overrightarrow{u} = \overrightarrow{v}(W')$ y $\overrightarrow{w} = \overrightarrow{v}(W^*)$ son vectores de contexto, véase la ecuación 2-2. Las probabilidades de palabras individuales $P(w_i)$ o probabilidades conjuntas de dos palabras $P(w_i, w_j)$, pueden ser estimadas con *Boolean document calculation*, en este caso, Cv incorpora una variante utilizando un cálculo de ventana

deslizante booleana, incluyendo así las frecuencias y distancias de las palabras, la cual se obtiene con la media aritmética de las medidas de confirmación individuales que se utilizan para llegar a una puntuación de coherencia del tema. ϕ es la medida de confirmación indirecta que calcula el aporte entre los subconjuntos de palabras de un par $S_i = (W', W^*)$, véase la ecuación 2-3, por tanto, $\phi_{S_i}(\overrightarrow{u}, \overrightarrow{w})$ se obtiene calculando la similitud del vector coseno entre todos los vectores de contexto $\overrightarrow{v}(W') \in \overrightarrow{u}$ y $\overrightarrow{v}(W^*) \in \overrightarrow{w}$ de un par $S_i = (W', W^*)$. S_i representa un par de subconjuntos de palabras, $S_i = (W', W^*)$ donde $W' \in W$, $W^* \in W$ y W consisten en las top-N palabras más probables del tema, definido formalmente como $S = \{(W', W^*)|W' = \{w_i\}; w_i \in W; W^* = W\}$. Por ejemplo, si $W = w_1, w_2, w_3$, un par posible sería $S_i = (W' = w_1), (W^* = w_1, w_2, w_3)$.

2.4. Trabajos relacionados

En las tablas **2-1** y **2-2** se sintetizan algunos trabajos relacionados que se consideraron más pertinentes por su naturaleza, metodología y aplicabilidad del modelado de temas en artículos científicos, como medio para comprender el área de investigación de un determinado conjunto de datos. Estos documentos se encuentran ordenados del más antiguo al más reciente, donde además de mostrar el año de publicación y los autores, en la tabla **2-1** se muestra de forma sintetizada el aporte metodológico o académico por el cual destaca el trabajo y una brecha que se identificó en dicho estudio que puede ser punto de partida para futuros trabajos. En la tabla **2-2** se hace una descripción de los métodos implementados y del conjunto de datos procesado de tal forma que, permita contextualizar el enfoque del método y su configuración de parámetros a un determinado ambiente de artículos científicos, esto es importante debido a que el modelado de temas y sus métodos son dependientes del contexto del conjunto de datos.

Por lo general en todos los trabajos se puede destacar que la configuración de parámetros como α, η y K (número de temas) para el método LDA son variables y dependen del contexto del conjunto de datos, como por ejemplo, variabilidad temática, número de documentos y grado de profundidad científica inciden en la asignación de valores y configuración del método de LDA. Por tal motivo, la exploración sistemática de parámetros del modelo es necesaria en este tipo de trabajos para hallar la mejor configuración según el conjunto de datos.

En la tabla **2-2** se hace un cita adicional en las celdas de la columna "Método"que permite identificar el tipo de método LDA implementado en algunos trabajos, dado que en [Blei et al., 2003] se proponen varios modelos, incluyendo uno base que en algunos trabajos fue implementado, y en [Blei, 2012] se establece el método concreto y mayormente utilizado, siendo el modelo usado en este trabajo, denominado "*smoothed LDA*", cuya denominación del método no es habitualmente usada refiriéndose indistintamente esta implementación como simplemente LDA.

Tabla 2-1.: Síntesis revisión de literatura. Aportes y brechas.

Año	Autor	Aporte	Brecha
2019	[Chen et al., 2019]	Metodología sistemática para el análisis de temas en literatura científica por medio de un análisis de modelos post-tema.	Rastrear el cambio y la evolución de temas temporales de forma automática.
2018	[Arrivillaga et al., 2016]	Aplicación interactiva que permite un análisis comparativo entre documentos y sus similitudes.	Incapaz de detectar cambios temáticos temporales.
2017	[Zhou et al., 2017]	Método de evolución de temas (citation-content-LDA) que toma en cuenta las relaciones de citación y contenidos de documentos.	Ausencia en la relación de autores y revistas en la evolución de los temas .
2017	[Syed and Spruit, 2017]	Estudio de coherencia de temas latentes a partir de resúmenes y textos completos.	Ausencia de análisis detallado entre cantidades similares de grandes conjuntos de datos.
2014	[Yau et al., 2014]	Estudio y aplicación de varios métodos de modelado de temas en un mismo conjunto de documentos científicos.	Permeabilidad en la evaluación de los resultados por un etiquetado manual por expertos.

Tabla 2-2.: Síntesis revisión de literatura. Descripción del conjunto de datos.

Año	Autor	Método	Conjunto de datos
2019	[Chen et al., 2019]	LDA base [Blei et al., 2003] (parámetros $\alpha = 0.5$ y $\beta = 0.01$, $K = 35$ temas).	Web of Science (Esto se asume dado que en el documento se hace mención a "World of Science - WoS"): 5,450 resúmenes (de 2,000 a 2,015), corpus de 15,585 términos. Dye-sensitized solar cell: 12,435 resúmenes (de 1991 a 2014), corpus de 41,214 términos.
2018	[Arrivillaga et al., 2016]	LDA base [Blei et al., 2003] ($K = 200$).	Bases de datos científicas.
2017	[Zhou et al., 2017]	citation-content-LDA (variante de LDA base [Blei et al., 2003]) ($K_{PAMI} = 30$ y $K_{CS} = 100$, $\alpha = 0.5$ y $\beta = 0.01$).	IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI): (de 1,995 a 2,012), 2,719 documentos, 6,284 citas y 886 términos únicos. IEEE Computer Society (CS) (de 1,967 a 2,006), 42,213 documentos, 33,961 citas y 5,704.
2017	[Syed and Spruit, 2017]	LDA [Blei, 2012] ($\alpha = (1/K)$ y $\eta = 1/V$ donde V es número de palabras).	Journal Canadian Journal of Fisheries and Aquatic Sciences: 4,417 documentos (de 1,996 a 2,016). 12 top-tier revistas de pesca (e. g. Canadian Journal of Fisheries and Aquatic Sciences y Fish and Fisheries): 15,004 documentos (de 2000 a 2,016).
2014	[Yau et al., 2014]	LDA base [Blei, 2012] $K = 50$ temas, CTM, Hierarchical LDA, HDP.	Web of Sciencie: 1,260 documentos clasificados en 7 clases para el proceso de prueba y 1,006 documentos para validación.

Según la literatura, existen múltiples retos en el modelado de temas en artículos científicos, algunos son la detección y análisis de relaciones de autores y revistas científicas [Zhou et al., 2017], el análisis de la causalidad de la dinámica en la evolución de los temas [Xiong et al., 2019], el descubrimiento de grupos por medio de grafos de temas [Jelodar et al., 2019], entre otros. Según [Arrivillaga et al., 2016], hay una creciente necesidad de técnicas automáticas para visualizar, analizar y resumir grandes colecciones de documentos implementando métodos de modelado de temas para el análisis y descubrimiento de temas, tendencias temporales e identificación de dinámicas de investigación, nuevas oportunidades y facilitar los procesos de gestión y generación de conocimiento de forma activa y dinámica [Arrivillaga et al., 2016, Alghamdi and Alfalqi, 2015]

3. Preparación y construcción del conjunto de datos

Este capítulo presenta el proceso metodológico realizado para la búsqueda y selección de artículos científicos. Posteriormente, se define el proceso metodológico para el preprocesamiento y preparación de los datos, en el cual se establecen las diferentes etapas implementadas. Seguidamente, se definen tres diferentes formas de representación textual. Por último, se describe el conjunto de datos consolidado.

3.1. Búsqueda y selección de artículos científicos

Se definió una metodología para el proceso de consolidación del conjunto de datos de artículos científicos. La figura **3-1** muestra la representación gráfica de dichos pasos los cuales se detallan a continuación:

1. Se aplicó la siguiente ecuación de búsqueda "AF-ID(60104325)" en SCOPUS[1], la cual es una ecuación de búsqueda clave para búsquedas avanzadas, y así encontrar artículos científicos publicados referentes a autores con filiación a la Universidad de los Llanos (también es posible realizar una búsqueda por nombre de la institución en la sección de "Afiliaciones" del buscador de SCOPUS).

2. Se seleccionaron todos los artículos científicos resultantes de la búsqueda disponibles. Al momento de conformar el conjunto de datos de este trabajo final, los artículos que se tomaron fueron desde el año 1999 hasta el 28 de octubre de 2019 (20 años).

3. Se extrajeron todos los metadatos disponibles en SCOPUS para cada uno de los artículos científicos y se consolidaron en un único archivo CSV donde cada fila corresponde a un registro de un artículo y cada columna a un metadato. Los 54 metadatos que se recopilaron directamente de SCOPUS fueron los siguientes: "ID", "Authors", "Author(s) ID", "Title", "Year", "Source title", "Volume", "Issue", "Art. No.", "Page start", "Page end", "Page count", "Cited by", "DOI", "Link", "Affiliations", "Authors with affiliations", "Abstract", "Author Keywords", "Index Keywords", "Molecular Sequence Numbers", "Chemicals/CAS", "Tradenames", "Manufacturers", "Funding Details", "Funding Text 1", "Funding Text 2", "Funding Text 3", "Funding Text 4",

[1]Enlace web: `https://www.scopus.com/search/form.uri`

"Funding Text 5", "Funding Text 6", "Funding Text 7", "References", "Correspondence Address", "Editors", "Sponsors", "Publisher", "Conference name", "Conference date", "Conference location", "Conference code", "ISSN", "ISBN", "CODEN", "PubMed ID", "Language of Original Document", "Abbreviated Source Title", "Document Type", "Publication Stage", "Access Type", "Source", "EID", "Scopus Subarea", "SciVal Topic Prominence".

4. Se obtuvieron los artículos científicos seleccionados en su formato PDF original, alojándolos en un directorio local, agregando así, su ruta relativa como un metadato extra en el archivo único CSV. Este campo se denominó "PDF Name" sumando un total de 55 columnas.

5. Se aplicó un proceso de transformación textual automático donde se extrajo todo el texto posible de los PDF's generando un archivo TXT por cada artículo científico. En este se trabajo se utilizó el módulo PDFMiner[2] de Python dado que, según revisión manual, fue la herramienta que mejor desempeñó tuvo, con menos errores de transformación en comparación con otras librerías, módulos o API's[3] disponibles usadas. Igualmente, todos los archivos TXT fueron alojados localmente.

6. El contenido textual de cada artículo se ingresó como una columna adicional al archivo único CSV. Es decir, que cada registro adicionó su contenido textual a los metadatos almacenados anteriormente. Dicho campo se nombró "FILE_TEXT" para un total de 56 columnas que conformaron el conjunto de datos base.

7. Se almacenó el archivo único CSV de manera local y en la nube por medio de repositorios web, como medio de respaldo. Así como también está disponible vía Kaggle[4].

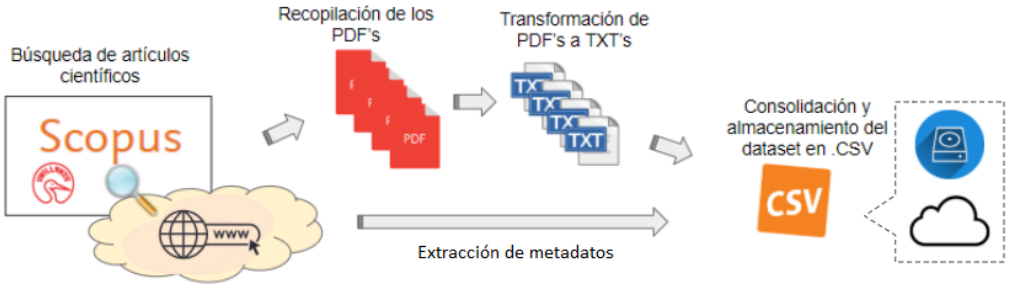


Figura 3-1.: Metodología para selección de artículos científicos. Elaboración propia.

[2] https://pypi.org/project/pdfminer/
[3] Interfaz de programación de aplicaciones (*Application programming interface*)
[4] Conjunto de datos disponible: DS_Unillanos_Papers_Full

Se utilizó el recurso electrónico SCOPUS tomando en cuenta que incluye el mayor solapamiento de las publicaciones científicas (editoriales y revistas) en Internet como por ejemplo RedALyC, SciELO, Web of Science o Google Scholar [Gavel and Iselid, 2008, Miguel, 2011, Jiménez Noblejas and Perianes Rodríguez, 2014, Martín-Martín et al., 2018].

3.2. Preprocesamiento - Preparación de los datos

En la figura 3-2 se representa gráficamente las etapas de la metodología propuesta para el preprocesamiento y preparación de los datos. Cada etapa generó un resultado intermedio que fue almacenado para garantizar respaldo y disminuir carga computacional. El resultado intermedio fue el insumo de la siguiente etapa.

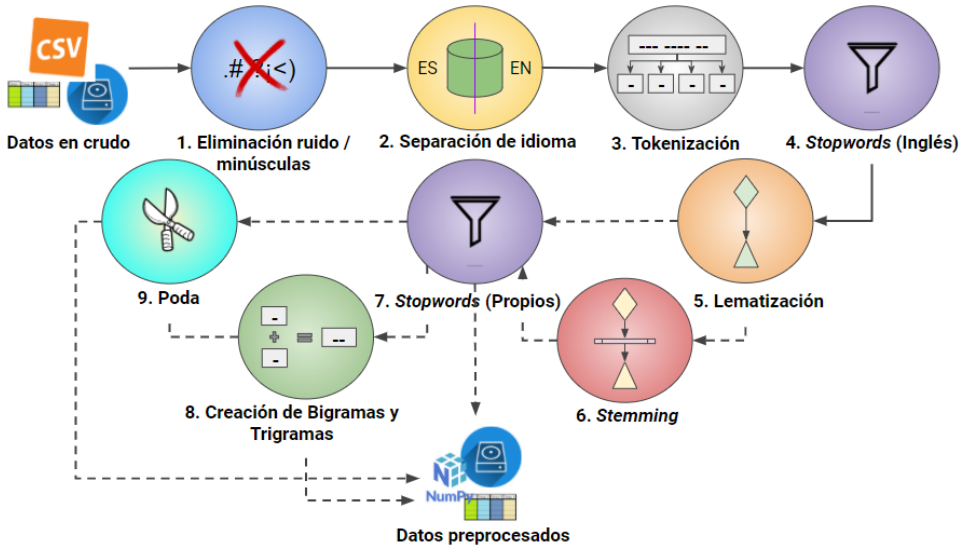


Figura 3-2.: Metodología de preprocesamiento del conjunto de datos. Elaboración propia.

A continuación, se detalla cada una de las etapas mostradas en la figura anterior las cuales consisten en:

1. **Eliminación de ruido/Transformación a minúsculas**: Se toma el archivo único CSV como fuente de datos para la eliminación de todos los caracteres no alfabéticos como números, símbolos y signos (e.g. ".", ",", "(", "2", "#", etcétera) exceptuando en esta etapa la eñe "ñ", las acentuaciones (tildes) "á", "é", "í", "ó" y "ú") y la diéresis "ü". Adicionalmente, se transforma todo el documento a minúscula para un manejo generalizado de las palabras.

2. **Separación de idioma**: Se divide el conjunto de datos seleccionando únicamente los artículos científicos escritos en inglés. Para ello, se usa el metadato "Language

of Original Document" que indica el idioma en que está escrito cada artículo. Los idiomas que se detectaron poseen mayor cantidad de documentos son inglés, español y portugués, siendo inglés el idioma más frecuente, razón por la cual se seleccionaron los artículos de éste idioma para este trabajo.

3. **Tokenización**: El contenido textual de cada artículo científico registrado en la columna "FILE_TEXT" se divide en sus palabras (tokens).

4. ***Stopwords* (Inglés)**: Se eliminan todas las palabras que no aportan un valor semántico al contenido textual, dichas palabras se denominan *stopwords*. Los *stopwords* aplicados en este caso son los más usados para el idioma inglés (e.g. "*of*", "*with*", "*the*", entre otros), a partir de, una lista predeterminada.

5. **Lematización**: Se lematiza cada palabra. Lematizar consiste en aplicar una transformación de palabras a su lema (término raíz) usando relaciones directas en un "diccionario" o base de datos predefinido [Heidenreich, 2018].

6. ***Stemming***: Consiste en aplicar un proceso de derivación (*stemming*) para transformar las palabras a su raíz semántica usando un algoritmo predefinido basado en la estructura de la palabra. En este caso se implementó el algoritmo de Porter [Porter, 1980].

7. ***Stopwords* (Propios)**: Se aplicó nuevamente un proceso de filtrado de *stopwords*, en este caso propios, creados a partir de una revisión manual de las palabras generadas hasta el momento. Estos *stopwords* propios fueron considerados a partir del contexto académico y científico de los artículos dado que existían palabras tanto técnicas como coloquiales que por su estructura o definición no aportaban valor semántico al conjunto de datos (e.g. siglas de enlaces químicos, "www", "https", abreviaturas como "com", "gov", "edu", números romanos como "xx" y "xv", palabras en español que existían en dichos documentos en inglés y algunas palabras en inglés que se consideraron eran *stopwords* propios del idioma y que no habían sido eliminados previamente de forma automática.

8. **Creación de Bigramas y Trigramas**: Se crean automáticamente términos cuyo significado está conformado por 2 o 3 palabras en conjunto (e.g. "small_shed").

9. **Poda**: Se aplica un criterio de poda para reducir el número de términos del conjunto de datos. Este criterio se detalla en la siguiente subsección.

Como se puede observar en la figura **3-2**, todas las etapas son secuenciales conectadas con un flecha de linea continua hasta la etapa 5 "Lematización". Posteriormente, el flujo de las etapas 6, 8 y 9 varía conectadas por flechas de linea puntuada, dado que se crearon todas las posibles combinaciones aplicando o no dichas etapas creando así un total de ocho conjuntos

de datos preprocesados diferentes. En otras palabras, un conjunto de datos preprocesado t_1 puede incluir el proceso de la etapa 6 "*Stemming*" y otro conjunto de datos preprocesado t_2 no, omitiendo este proceso. Esto fue propuesto para realizar una búsqueda sistemática de parámetros óptimos que permitiera escoger la mejor configuración posible para la representación textual de los artículos científicos, en este caso, en la fase de preprocesamiento y preparación de los datos. Un resumen de las variantes se muestra en la tabla **3-1**. El conjunto de datos base hasta culminada la etapa 2 "Separación de idioma" y referente a documentos en Inglés, está disponible en Kaggle[5]. Igualmente, otra versión del mismo habiendo culminada la etapa 5 "Lemmatización" tabién está disponible en Kaggle[6] en formato NPY (Numpy).

Tabla 3-1.: Etapas de preprocesamiento y su aplicación en la metodología propuesta.

Preprocesamiento	
Etapa	**Aplica**
1. Eliminación de ruido/Transformación a minúsculas	Sí
2. Separación de idioma	Sí
3. Tokenización	Sí
4. *Stopwords* (Inglés)	Sí
5. Lematización	Sí
6. *Stemming*	Sí/No
7. *Stopwords* (Propios)	Sí
8. Creación de Bigramas y Trigramas	Sí/No
9. Poda	Sí/No

En este trabajo de utilizaron los siguientes módulos de Python para el preprocesamiento de los datos: REGEX[7] en la etapa 1, Gensim[8] [Rehurek and Sojka, 2010] en las etapas 3, 4, 7 y 8, NLTK[9] en las etapas 5 y 6; y por último, Numpy[10] y Pandas[11] fueron implementados en prácticamente todas las etapas para el manejo y respaldo de los datos, tanto, de los resultados intermedios como de los finales.

3.2.1. Criterio de Poda

La Ley de Zipf es una ley empírica que señala una tendencia logarítmica descendente en las frecuencias de los términos usados en un idioma del lenguaje natural, donde los términos ordenados descendentemente por su frecuencia de aparición son clasificados en un ranking donde el primer término, siendo el más frecuente, ocupa el primer lugar con una frecuencia

[5]Conjunto de datos disponible: DS_Unillanos_Papers_EN
[6]Conjunto de datos disponible: DS_Unillanos_UpToLemm
[7]Enlace: `https://pypi.org/project/regex/`
[8]Enlace: `https://radimrehurek.com/gensim/`
[9]Enlace: `https://www.nltk.org/`
[10]Enlace: `https://numpy.org/`
[11]Enlace: `https://pandas.pydata.org/`

n y el segundo término, el segundo más frecuente, con una frecuencia igual a $2/n$, generando así que las frecuencias para los términos posteriores serían iguales a r/n donde r corresponde al r-ésimo lugar del término en el ranking, véase la ecuación 3-1. Este comportamiento en un plano bidimensional logarítmico en ambos ejes da como resultado una recta descendente [Zipf, 1949]. La figura **3-3** permite visualizar el comportamiento de la Ley de Zipf tanto en el plano de frecuencias (izquierda) como logarítmico (derecha). Existe una variante a la ecuación 3-1 que permite "suavizar" la recta para se ajuste mejor a los datos, para ello se agregaron dos constantes α y C que son calculadas de acuerdo al conjunto de términos, siendo que $\alpha \approx 1$, véase la ecuación 3-2. La Ley de Zipf ha podido representar múltiples idiomas del lenguaje natural (e.g. Español, Inglés, Francés, entre otros) [Piantadosi, 2014]. Adicionalmente ha sido ampliamente usada como punto de referencia para procesos computacionales que procesan el lenguaje natural [De Battisti et al., 2015, Srinivasa-Desikan, 2018].

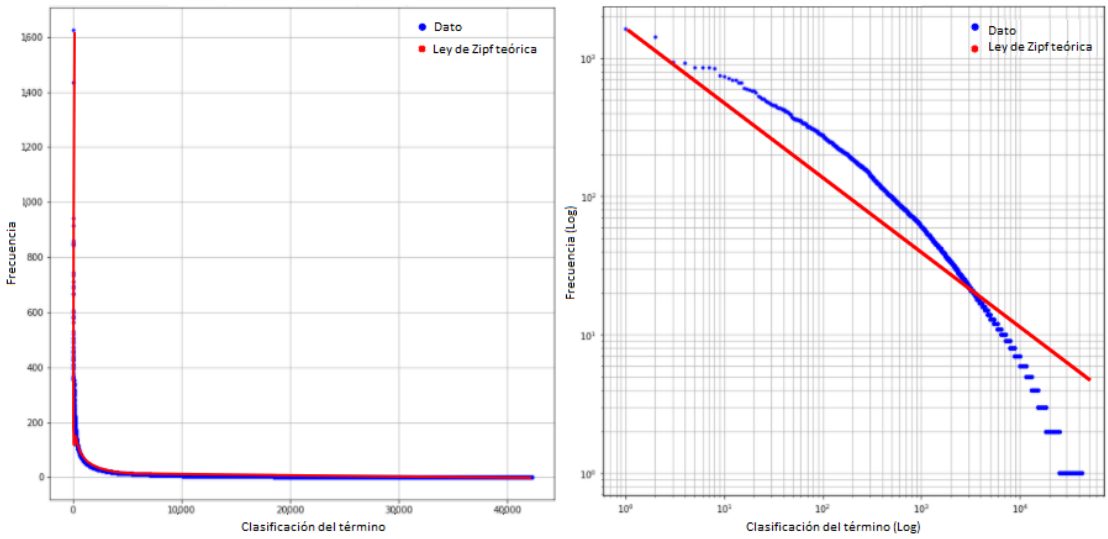


Figura 3-3.: Ley de Zipf teórica (rojo) y frecuencia de términos ordenados de un conjunto de datos (azul). Izquierda: Plano de frecuencias. Derecha: Plano logarítmico en ambos ejes. Elaboración propia.

$$f(r) = \frac{1}{r} \tag{3-1}$$

$$f(r) = \frac{C}{r^{\alpha}} \tag{3-2}$$

Posteriormente, Mandelbrot propuso una generalización de la variante de la Ley de Zipf que permitiera a la función ser más "flexible" y ajustarse mejor a los datos. Para ello agregó una

nueva constante $\beta \approx 2{,}7$ que puede ser ajustada según los datos, al igual que las otras dos constantes, véase la ecuación 3-3 [Mandelbrot, 1953].

$$f(r) = \frac{C}{(r+\beta)^{\alpha}} \tag{3-3}$$

Tomando en cuenta ambas propuestas, se hizo un proceso comparativo entre ellas para escoger la función con mejor ajuste a los datos. Para ello se tomó varios conjuntos de datos resultantes de las etapas anteriores a la etapa "Poda" con configuraciones distintas y, en cada experimento, la ecuación de Mandelbrot obtuvo mejores resultados. Por tal motivo, se toma la generalización de Mandelbrot como función idónea para representar un comportamiento cercano del conjunto de datos al lenguaje natural y así aplicar un proceso de eliminación de términos que no cumplan con tal criterio. Para ello, se propuso descartar los términos que presentaran una distancia mayor o igual a un umbral establecido. La distancia es calculada entre el punto real del dato y el punto teórico idóneo en el plano logarítmico, véase la figura **3-4**. El umbral establecido fue de 0.35 según criterio propio de acuerdo con lo observado en los diferentes conjuntos de datos.

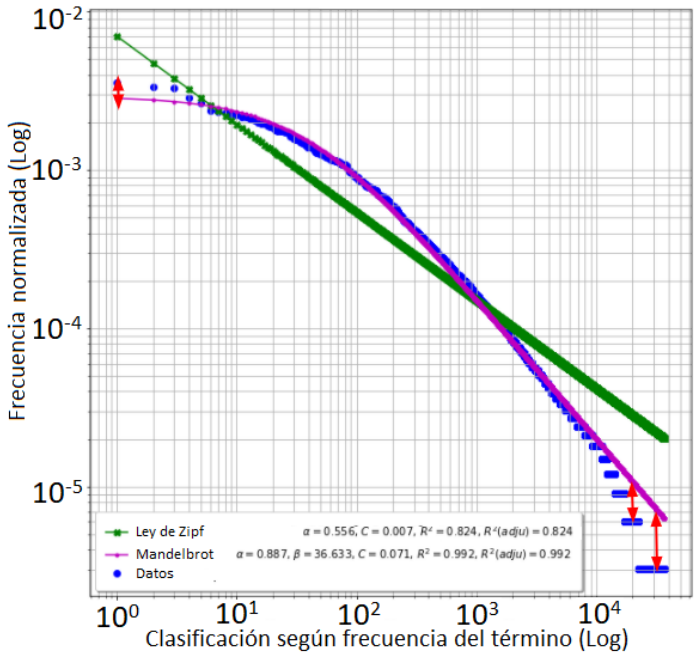


Figura 3-4.: Visualización de la distancias tomadas entre los datos reales y la función idónea (Mandelbrot) en el plano logarítmico. Elaboración propia.

3.3. Representación textual

Los datos de tipo textual son datos no estructurados considerados de gran dimensionalidad, dado que cada término representa una dimensión. Por tal motivo, existen modelos computacionales propuestos que permiten representar relaciones textuales en un espacio dimensional, ejemplos de estos modelos son: el modelo booleano, el modelo espacio vectorial (*Vector Space Model*, VSM), el análisis semántico latente (LSA), los grafos, entre otros [Torres-López and Arco-García, 2016]. VSM es uno de los modelos de representación textual más usados para realizar tareas de recuperación de información y NLP. VSM representa documentos textuales a través de vectores de términos, un ejemplo de esta representación vectorial se puede visualizar en la figura **3-5** donde cada término representa una dimensión (en este caso el espacio es tridimensional) y cada documento D del conjunto de datos se expresa como un vector $\overrightarrow{D}$ de tamaño W donde W es la cantidad de términos únicos en el conjuntos de datos y cada posición del vector corresponderá al valor del término w para el documento d. El vector $\overrightarrow{Q}$ es el documento Q a comparar. En esta representación, el documento más similar a Q será el que tenga un menor ángulo θ con él. VSM puede ser expresado por diferentes tipos de modelos distribucionales basados en vectores de conteo, donde el valor de conteo corresponde a la frecuencia de aparición de unidades (términos) en un documento o texto dado. Este conteo se transforma en una matriz, cuya estructura varía según el enfoque. En términos generales, existen tres enfoques: similitud de documentos (matrices término-documento), similitud de palabras (matrices palabra-contexto) y similitud de relaciones (matrices par-patrón) [Torres-López and Arco-García, 2016].

Se crearon tres representaciones textuales por cada conjunto de datos preprocesado. Esto se hizo con el objetivo de encontrar el mejor modelo de representación textual para el conjunto de datos. En esta parte se crearon 24 modelos diferentes en combinación con la etapa anterior. Las representaciones textuales usadas fueron: Bolsa de palabras (*Bag of Words* - BoW), Frecuencia de términos - Frecuencia inversa de documentos (*Term frequency - Inverse document frequency*, TF-IDF) y Matriz binaria o Matriz de incidencia.

3.3.1. Bolsa de palabras (Bag of Words - BoW)

Consiste en un vector de documentos donde en cada posición, que corresponde a un término basado en un diccionario único de términos, se asigna su frecuencia de aparición, a esto se le denomina Frecuencia de términos (o *Term Frequency*, tf): $tf(w, d)$, donde tf es la frecuencia del término (cantidad de ocurrencias del término w en un documento d). La BoW corresponde a un tipo de matriz término-documento. Tomando como ejemplo la tabla **3-2**, es posible observar que la columna "Términos" hace referencia a un diccionario de términos únicos ordenados alfabéticamente cuyo tamaño W corresponde a la cantidad de términos

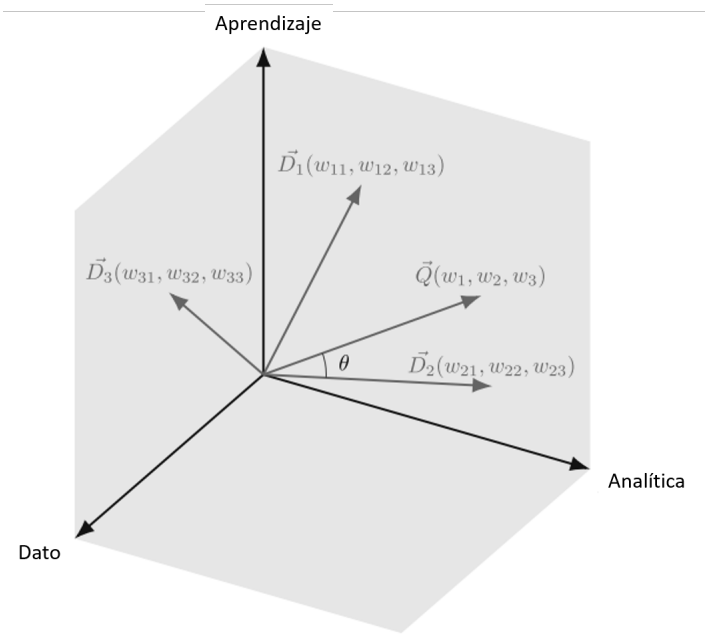


Figura 3-5.: Visualización VSM en un espacio de tres dimensiones (términos), correspondientes a {Dato, Analítica, Aprendizaje}. Adaptado de [Gudivada et al., 2018].

únicos de todo el conjunto de datos, siendo w_1 el primer término ("artificial") y w_T el w-ésimo término ("zafiro") [Torres-López and Arco-García, 2016, Gudivada et al., 2018]

Tabla 3-2.: Ejemplo Matriz término-documento. Representación BoW.

	Documentos						
Términos	d_1	d_2	d_3	d_4	d_5	...	d_D
artificial	0	2	89	1	0	...	0
computador	1	1	0	1	0	...	0
inteligencia	0	20	0	1	0	...	0
investigación	0	0	0	1	70	...	0
...	...	...	...	...	...	...	...
zafiro	0	1	0	24	0	...	0

La BoW ignora el orden en que los términos aparecen en los documentos por lo que se ignora completamente la estructura lingüística del texto. Como se puede observar en la tabla **3-2** la mayoría de elementos de esta matriz son 0, a esto se le denomina, matriz dispersa, dado que la mayoría de documentos contienen únicamente una parte de todo el diccionario. Según las matrices término-documento son similares los documentos cuando sus vectores columna son similares. Adicionalmente, este tipo de matrices son muy usadas en el área de recuperación de información [Torres-López and Arco-García, 2016].

3.3.2. Frecuencia de términos - Frecuencia inversa de documentos (Term frequency – Inverse document frequency, TF-IDF)

TF-IDF es una matriz término-documento que permite expresar un peso relativo del término w en el vector asociado a un documento d considerando la cantidad de documentos que poseen dicho término [Torres-López and Arco-García, 2016]. La expresión matemática de TF-IDF está conformada por las ecuaciones 3-4 y 3-5.

$$tfidf(w,d) = tf(w,d) * idf(w) \tag{3-4}$$

$$idf(w) = \log \frac{N}{df(w)} \tag{3-5}$$

donde $tf(w,d)$ es la frecuencia del término w en el documento d, $idf(w)$ es la frecuencia inversa de documentos, es decir, el número de documentos donde existe el término w pero de forma inversa, esto con el objetivo de asignar más peso a los términos que existen en una menor cantidad de documentos, N es el número de documentos en el conjunto de datos, y $df(w)$ es la frecuencia de documentos, es decir, el número de documentos donde el término w existe [Torres-López and Arco-García, 2016, Gudivada et al., 2018]. Tomando el ejemplo mostrado en la tabla **3-2**, la matriz término-documento usando TF-IDF sería algo similar a la tabla **3-3**.

Tabla 3-3.: Ejemplo Matriz término-documento. Representación TF-IDF.

	Documentos						
Términos	d_1	d_2	d_3	d_4	d_5	...	d_D
artificial	0.011	0.002	0.032	0.002	0.096	...	0.051
computador	0.065	0.080	0.042	0.050	0.028	...	0.030
inteligencia	0.015	0.087	0.048	0.032	0.027	...	0.019
investigación	0.010	0.036	0.062	0.088	0.011	...	0.018
...	...	...	...	...	...	...	...
zafiro	0.022	0.034	0.037	0.039	0.032	...	0.045

3.3.3. Matriz binaria o Matriz de incidencia

El modelo booleano genera una matriz término-documento cuyo valor es booleano o binario (0 ó *false* para la ausencia y 1 ó *true* para la existencia). Por tal motivo, a este modelo también se le conoce como matriz binaria, booleana o de incidencia. Esta representación es más rápida en términos de tiempo computacional por que requiere menos espacio en memoria al tener únicamente valores binarios, aunque en muchos casos no es más eficiente dado que

no permite generar pesos ponderados como en TF-IDF o relacionar frecuencias como en BoW [Torres-López and Arco-García, 2016, Gudivada et al., 2018]. Un ejemplo de la matriz término-documento se muestra en la tabla **3-4**.

Tabla 3-4.: Ejemplo Matriz término-documento. Representación Binaria.

	Documentos						
Términos	d_1	d_2	d_3	d_4	d_5	...	d_D
artificial	0	1	1	1	0	...	0
computador	1	1	0	1	0	...	0
inteligencia	0	1	0	1	0	...	0
investigación	0	0	0	1	1	...	0
...	...	...	...	...	...	...	...
zafiro	0	1	0	1	0	...	0

3.4. Descripción conjunto de datos

El conjunto de datos consolidado en el archivo único CSV contiene 293 filas donde la primera fila contiene los encabezados y el resto de filas corresponden cada una a un artículo científico y 56 columnas que fueron mencionadas en la sección 3.1, de los cuales caben resaltar y detallar las mostradas en la tabla **3-5**. En la misma tabla se puede observar dos artículos científicos como muestra del contenido en algunas columnas.

La columnas "Scopus Subarea" y "SciVal Topic Prominence"[12] son etiquetas asignadas automáticamente por SCOPUS a cada artículo científico, donde "Scopus Subarea" indica grupos más generales, denominados áreas, como "Computer Science" y "SciVal Topic Prominence" grupos más específicos como "Histopathological images". Cada artículo puede tener múltiples etiquetas las cuales son separadas por el caracter "|". Las cinco áreas, de un total de 26, con mayor número de documentos asignadas son "Agricultural and Biological Sciences"(140), "Veterinary" (56), "Engineering" (32), "Physics and Astronomy" (26) y "Computer Science" (25).

La columna "Language of Original Document" fue usada como indicador para filtrar artículos científicos por idioma permitiendo seleccionar 137 documentos en Inglés. Su distribución por áreas de SCOPUS no varió mucho y las cinco áreas de mayor densidad de documentos son "Agricultural and Biological Sciences"(57), "Engineering"(24), "Computer Science"(22), "Physics and Astronomy"(21) y "Environmental Science"(17), véase la figura **3-6**.

[12]Fuente Elsevier: https://www.elsevier.com/solutions/scival/releases/topic-prominence-in-science

Tabla 3-5.: Muestra del conjunto de datos. Cuatro de 56 columnas.

Columna	**Descripción**	**Muestra dato 1** Porras-García et al., 2018	**Muestra dato 2** Cano et al., 2018
Language of Original Document	Idioma en que se encuentra escrito el artículo, puede ser un único idioma o combinado.	Spanish	English
Scopus Subarea	Subáreas al que pertenece determinado artículo, según Scopus.	Computer Science | Decision Sciences	Engineering | Physics and Astronomy | Computer Science | Mathematics | Materials Science
SciVal Topic Prominence	Agrupación de publicaciones en temas hecha por Scopus basado en un análisis de citas directas.	Ontology | Semantics | Legal knowledge	Medical imaging | Pathology | Histopathological images
FILE_TEXT	Contenido textual extraído automáticamente del artículo científico.	*"...Este trabajo presenta un ana´lisis comparativo de la implementacio´n de una representacio´n de Bolsa de Palabras (Bag of Words - BoW) para el procesamiento distribuido de una coleccio´n de documentos de texto en la plataforma de proce- samiento distribuido Apache SparkTM..."*	*"...The Convolutional neural networks (CNN) have been shown to be able to learn the relevant visual features for different computer vision tasks from large amounts of annotated data. Hence, the performance of CNNs can vary depending on the training data set and associated model architecture...."*

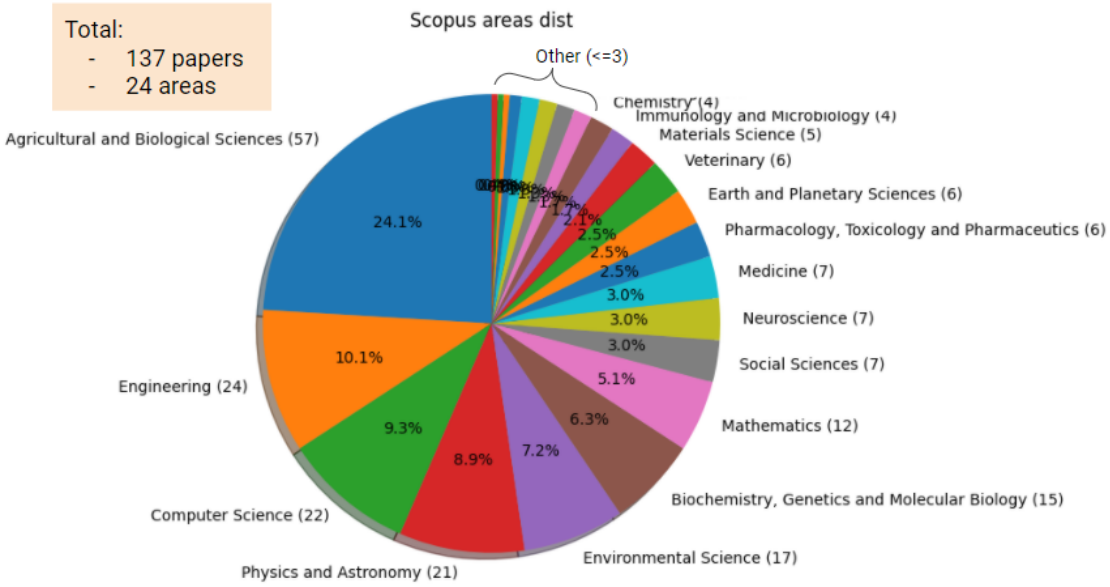


Figura 3-6.: Distribución de 137 artículos científicos en inglés por áreas según SCOPUS. Elaboración propia.

4. Modelado y análisis de temas de artículos científicos

Este capítulo presenta la Asignación Latente de Dirichlet (*Latent Dirichlet Allocation - LDA*) como método para la aplicación del modelado de temas. Seguido, se describe todo el proceso de modelado y parametrización LDA por medio de una búsqueda sistemática de parámetros. Posteriormente, se abordan los métodos implementados para la exploración y visualización de temas latentes obtenidos por LDA. Por último, se detalla el diseño experimental de evaluación cuantitativa y cualitativa.

4.1. Asignación Latente de Dirichlet (Latent Dirichlet Allocation - LDA)

LDA es un modelo generativo probabilístico bayesiano de temas que asume que los documentos se representan como mezclas de temas latentes aleatorios y que cada tema posee una distribución de probabilidad de palabras [Alghamdi and Alfalqi, 2015]. En el modelo gráfico de la figura **4-1** se pueden visualizar las representaciones de las variables aleatorias ocultas (círculos blancos) como las distribuciones documento-tema, palabra-tema y tema-palabra, así como, la variable aleatoria visible (círculo negro) de la distribución de palabras observadas, de acuerdo con la representación del flujo inverso de escritura del ser humano según [Blei et al., 2003]. Por un lado, la definición de la probabilidad condicional para la asignación de palabras a temas está definida por las ecuaciones 4-1 y 4-2 [Blei et al., 2003, Blei, 2012] que son basadas en el teorema de Bayes [Bayes, 1763]. Siendo α y η los hiperparámetros de las distribuciones Dirichlet respectivas, α y $\eta \in \mathbb{R}^+$; K es el número de temas $[1, K]$; D el número de documentos $[1, D]$; N es el número de palabras $[1, N]$, en el documento $d \in D[1, Nd]$; θ_d es la distribución documento-tema, $d \in D$; β_k corresponde a la distribución tema-corpus, $k \in K$; $z_{d,n}$ es la asignación tema-palabra para un documento $d \in D$; $w_{d,n}$ son las palabras observadas y no ocultas para un documento $d \in D$. La ecuación 4-1 consiste en hallar la distribución de probabilidad conjunta de las distribuciones $\beta_{1:K}$, $\theta_{1:D}$, $z_{1:D}$ y $w_{1:D}$ dado los hiperparámetros α y η, para lo cual se hallan cada una de las distribuciones de manera individual y posteriormente, se obtiene la probabilidad conjunta partiendo de la suposición bayesiana, de la independencia de cada distribución, por lo cual se multiplican

las distribuciones. Respecto a la ecuación 4-2 es donde directamente se aplica el teorema de Bayes donde a partir de las distribuciones de probabilidad conjuntas (numerador) dividido por la distribución de probabilidad de los términos se obtiene la probabilidad condicional de las distribuciones dados los términos. Se aplica de esta forma siendo que a priori la única distribución que se conoce es de los términos del conjunto de datos para así obtener a posteriori las distribuciones de probabilidad dada dicha distribución.

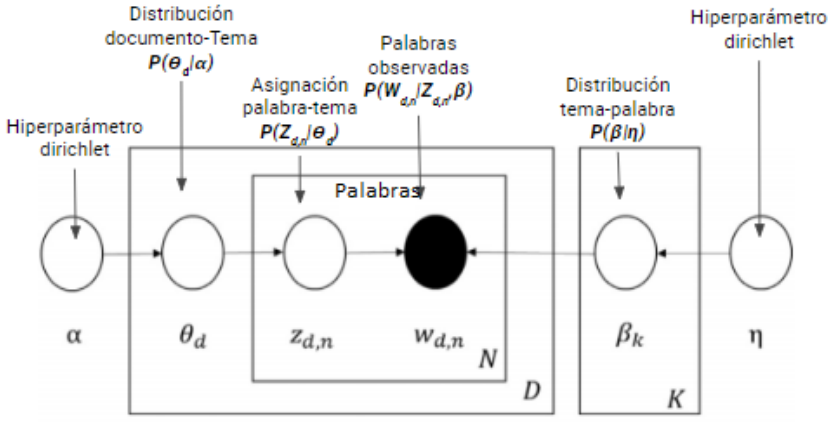


Figura 4-1.: Modelo gráfico de LDA. Adaptado de [Allahyari et al., 2017, Blei, 2012].

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}|\alpha, \eta) = \prod_{k=1}^{K} p(\beta_k|\alpha, \eta) \prod_{d=1}^{D} p(\theta_d|\alpha, \eta) \left(\prod_{n=1}^{N} p(z_{d,n}|\theta_d) p(w_{d,n}|\beta_{1:K}, z_{d,n}) \right) \tag{4-1}$$

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}|w_{1:D}, \alpha, \eta) = \frac{p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}|\alpha, \eta)}{p(w_{1:D}|\alpha, \eta)} \tag{4-2}$$

El método de LDA se describe en el algoritmo 1. Donde primero se asigna una distribución Dirichlet para η por cada tema k y así construir la distribución β_k. Posteriormente, para cada documento d se define una distribución de temas a partir de una distribución Dirichlet de α (θ_d) y para cada término w_n de dicho documento d se asigna un tema z_n a partir una distribución multinomial de la distribución θ_d y se muestrea un término w_n aplicando una distribución multinomial de β_{z_n}. En el algoritmo se estable como "palabra" al concepto o unidad que conforma los documentos y en este estudio se usa el concepto de "término", esto se debe a dos aspectos: se estableció una metodología para construir términos de bigramas o trigramas para el diccionario; y que, por fines prácticos y según lo encontrado en la literatura, el algoritmo definido habitualmente usa este concepto.

Algoritmo 1 Algoritmo LDA

para tema $k = 1, 2, \ldots, K$ **hacer**
    muestrear una distribución de palabras $\beta_k \sim Dir(\eta)$
fin para
para documento $d = 1, 2, \ldots, D$ **hacer**
    Muestrear una distribución de temas $\theta_d \sim Dir(\alpha)$
    para palabra w_n, $n = 1, 2, \ldots, N$, en el documento d **hacer**
        Muestrear un tema $z_n \sim Mult(\theta_d)$
        Muestrear una palabra $w_n \sim Mult(\beta_{z_n})$
    fin para
fin para

4.2. Modelado y parametrización del algoritmo LDA

La implementación del método LDA fue realizada en Python usando Gensim [Rehurek and Sojka, 2010]. Cabe aclarar que Gensim implementa internamente una variación de LDA (versión "*smoothed*" y generalizada en la literatura como LDA) denominada Online LDA, presentada por [Hoffman et al., 2010], la cual es una implementación del algoritmo de Bayes variacional en línea (*Online Variational Bayes*) para LDA basada en la optimización estocástica en línea con un paso de gradiente natural, que converge a un óptimo local de la función objetivo del algoritmo de Bayes variacional. Por ser una versión en *online* del algoritmo, permitirá actualizar el modelo en el tiempo a partir del modelo actual en la medida de que el conjunto de documentos de artículos científicos aumente proporcional o masivamente cada año [Rehurek and Sojka, 2010, Hoffman et al., 2010]. Esta implementación es la más usada en el área de NLP para tareas como TM [Jelodar et al., 2019]. Además de los parámetros requeridos por el método LDA, como son el número de temas K y los hiperparámetros α y η, Gensim implementa parámetros adicionales que permiten ajustar el método LDA ya sea manual o automáticamente de acuerdo con la representación textual y el conjunto de datos.

Por tal razón, se realizó una Búsqueda Sistemática de Parámetros (Systematic Parameter Search - SPS) específicamente para la implementación LDA con Gensim con el objetivo de analizar e identificar el impacto y sensibilidad de los parámetros en el desempeño del modelo entrenado, para usar la mejor configuración de parámetros para el modelo de acuerdo al conjunto de datos. Los parámetros de mayor impacto para su exploración y análisis, de acuerdo con la implementación en Gensim fueron:

- **alpha**: Hiperparámetro α de LDA. Pueder ser un vector de tamaño K, donde cada posición i corresponderá al peso a priori del tema i, o una de las siguientes cadenas de texto: *i)* "symmetric", la cual aplica una distribución uniforme con valor $1/K$ a cada tema, *ii)* "asymmetric", que asigna un peso de $1/k$ donde k es el número del tema, y *iii)*

"auto"la cual utiliza métodos de inferencia automáticos para detectar la distribución de pesos óptima a partir de los datos. Específicamente, Gensim implementa *Online Variational Bayes* [Hoffman et al., 2010, Syed and Spruit, 2018].

- **eta**: Hiperparámetro η de LDA. Puede ser un valor escalar para una probabilidad simétrica a priori sobre tema/palabra, un vector de tamaño W para denotar una probabilidad asimétrica definida manualmente para cada palabra, donde W es el número de términos, o bien una matriz de forma (K, T) para asignar una probabilidad para cada combinación palabra-tema, o una cadena de texto: *i)* "symmetric" con una distribución uniforme con valor $1/T$, o *ii)* "auto" que calcula pesos de igual forma que **alpha**.
- "**num_topics**": Número de temas latentes a identificar (K).
- "**chunksize**": Cantidad de documentos a procesar al mismo tiempo por el método LDA durante el entrenamiento. Un mayor número de documentos permite reducir el tiempo computacional requerido, siempre y cuando dicha cantidad se pueda alojar en la memoria RAM del equipo de cómputo [Srinivasa-Desikan, 2018].
- "**passes**": Frecuencia de entrenamiento del modelo en todo el conjunto de datos. También se puede denominar como épocas [Srinivasa-Desikan, 2018].
- "**update_every**" ó "**batch**": Número de documentos que se van a iterar para cada actualización. Establecido a 0 para el aprendizaje por lotes (*batch*), > 1 para el aprendizaje iterativo en línea (*online LDA*).
- "**decay**": Hace referencia al valor κ de [Hoffman et al., 2010]. En términos generales, es un número entre $(0, 5$ y $1]$ para ponderar qué porcentaje del valor λ, en [Hoffman et al., 2010], se olvida cuando se examina cada nuevo documento [Rehurek and Sojka, 2010].
- "**iterations**": Frecuencia de repetición de un bucle determinado en cada documento. Es importante establecer valores altos para "**passes**" e "**iterations**" [Srinivasa-Desikan, 2018].

La tabla **4-1** permite ver los posibles valores a evaluar para cada uno de los parámetros. Todos los valores de los parámetros fueron basados en argumentos como la documentación propia de la herramienta y sus valores por defecto[1], trabajos similares [Hoffman et al., 2010, Syed and Spruit, 2018, Srinivasa-Desikan, 2018] y criterio propio según el contexto y cantidad de artículos científicos. La SPS consistió en crear un modelo LDA entrenado por cada combinación posible de parámetros variando su rango u opciones, a partir de un conjunto de datos de un tipo de representación textual. En total se entrenaron 103,680 modelos con configuraciones diferentes, 4,320 modelos por cada representación textual.

[1]Enlace: `https://radimrehurek.com/gensim/models/ldamodel.html`

Tabla 4-1.: Descripción de valores de exploración de parámetros del método LDA con Gensim.

Procesamiento		
Parámetro	**Aplicación**	**Justificación**
"**alpha**"	["symmetric"; "asymmetric"; "auto"]	La asignación manual de valores es dependiente del conocimiento de la distribución de los temas previamente. Por lo que se usó la configuración que permitiera a la herramienta automatizar dicho proceso.
"**eta**"	["symmetric"; "auto"]	
"**num_topics**"	[10; 20; 30; 50]	Tomando en cuenta la cantidad de documentos (i.e. 137 artículos científicos).
"**chunksize**"	[10; 20; 30; 50; 137]	Comparar el uso de recursos computacionales usando cantidades de documentos diferentes.
"**passes**"	[1; 7; 15]	Basado en el costo computacional.
"**update_evary**" ó "**batch**"	[0; 1]	Escoger una implementación de LDA: LDA convencional (*batch*) [Blei et al., 2003] u *Online LDA* [Hoffman et al., 2010].
"**decay**"	[0.51; 1]	Los valores extremos permitidos dado el alcance del trabajo final.
"**iterations**"	[25; 50; 70]	Cantidades proporcionales al número de documentos.

4.3. Exploración y visualización de temas

La exploración y visualización de temas es uno de los retos actuales existentes en el modelado de temas [Jelodar et al., 2019]. Existen herramientas como PyLDAvis que permiten representar los temas resultantes del modelado de temas en un campo visual bidimensional [Sievert and Shirley, 2014]. Esto permite identificar relaciones de semejanza o cercanía entre temas por medio de una comparación visual, esto favorece la comprensión y análisis por medio de representaciones gráficas que facilitan su interpretación, siendo esto uno de los principales problemas en el modelado de temas y por el cual cada trabajo propone implementaciones o herramientas acorde al enfoque de sus estudios [Jelodar et al., 2019]. En este trabajo se usó PyLDAvis para la visualización de temas dado que es la herramienta más usada para abordar dicho problema según la literatura revisada [Sievert and Shirley, 2014, Shiryaev et al., 2017, Syed and Spruit, 2017, Syed and Spruit, 2018]. Adicionalmente, se usó el algoritmo t-distributed Stochastic Neighbor Embedding (t-SNE) para la visualización de documentos y temas latentes dominantes, gracias a que permite visualizar conjuntos de datos de alta dimensionaliad, siendo éste el caso de la MT.

4.3.1. PyLDAvis

PyLDAvis[2] es un módulo de Python para la visualización de temas latentes generados por TM. Es una implementación en Python del trabajo propuesto por Carson Sievert en [Sievert and Shirley, 2014], cuya representación ha sido ampliamente usada para la visualización de TM [Syed and Spruit, 2017, Shiryaev et al., 2017, Syed and Spruit, 2018]. PyLDAvis sirve para generar una representación visual que permita ayudar a interpretar los resultados de la aplicación de un modelo entrenado de TM. PyLDAvis extrae información del modelo que es usada para generar una visualización interactiva. Esta visualización consiste básicamente en dos elementos: el lado izquierdo compone la visualización bidimensional que muestra la distancia entre temas por medio del escalamiento multidimensional (Multidimensional scaling, MDS) y el lado derecho muestra las relaciones de los términos para cada uno de los temas por medio de histogramas de frecuencias [Sievert and Shirley, 2014]. Según [Guerrero-Casas, 2012], MDS en términos generales, es una "*técnica multivariante de interdependencia que trata de representar en un espacio geométrico de pocas dimensiones las proximidades existentes entre un conjunto de objetos o de estímulos*". Existen múltiples tipos de MDS, pero en esencia, todas toman una matriz, por lo general, de proximidades (distancias o similitudes) proveniente del conjunto de objetos y la transforman en una matriz de coordenadas donde la cantidad de columnas dependerá de las dimensiones establecidas. MDS sirve para representar visualmente los objetos y así determinar relaciones entre ellos [Guerrero-Casas, 2012]

4.3.2. t-distributed Stochastic Neighbor Embedding, t-SNE

La incrustación estocástica de vecinos distribuida en t (t-distributed Stochastic Neighbor Embedding, t-SNE) es un algoritmo diseñado para la visualización de conjuntos de datos de alta dimensionalidad [van der Maaten and Hinton, 2008]. t-SNE realiza dos pasos que son explicados de forma general:

1. Construye una distribución de probabilidad de parejas de muestras en el espacio inicial (original). Tal que las muestras semejantes reciben una probabilidad mayor de ser seleccionadas, a diferencia de las muestras diferentes con una probabilidad mucho menor.

2. Aplica un proceso de reducción de la dimensionalidad llevando los puntos a un espacio con baja dimensionalidad aleatoriamente, por lo cual define una distribución de probabilidad semejante a la construida en el espacio inicial [Interactive Chaos, 2020].

[2]Enlace: `https://pyldavis.readthedocs.io/en/latest/`

4.4. Diseño experimental

4.4.1. Evaluación cuantitativa

Se planteó una Búsqueda Sistemática de Parámetros (SPS, por sus siglas en Inglés) general que permitiera articular las fases anteriores ("Preprocesamiento" y "Representación textual" en las secciones 3.2 y 3.3, respectivamente) con el modelado de temas usando LDA y posteriormente realizar un proceso de evaluación de desempeño con la medida de Coherencia (sección 2.3) para cada uno de los modelos entrenados. Para ello, se siguieron los siguientes pasos:

1. Por cada cada conjunto de datos preprocesado se generaron tres representaciones textuales (i.e. BoW, TF-IDF y Matriz Binaria).

2. Por cada representación textual generada se entrenaron 4,320 modelos LDA correspondientes a las variaciones posibles de parámetros mostradas en la tabla **4-1**.

3. Por cada modelo LDA entrenado se realizó un proceso de evaluación de desempeño usando cuatro diferentes CM variando el tamaño de ventana deslizante en cada una. Los posibles valores fueron 10, 25, 50 y 110. Esto se realizó para observar el impacto del tamaño de la ventana deslizante en los conjuntos de datos y determinar un tamaño óptimo para la CM.

4. Se analizaron todas las medidas de desempeño de las CM de manera conjunta y se seleccionaron los mejores modelos, es decir, aquellos con mayor CM, por cada tamaño de ventana deslizante. De los cuales, se seleccionó el modelo LDA entrenado con mayor CM en general y se definió su configuración de parámetros como combinación base.

5. Para el análisis y criterio de selección de cada etapa o parámetro en la SPS, se optó por usar la configuración de parámetros base seleccionada en el paso anterior y se graficó usando como variable independiente en el eje "x" el parámetro del número de temas variando únicamente dicha etapa o parámetro a analizar como variable dependiente en el eje "y". Esto, para observar el impacto de cada parámetro a medida que el número de temas asignado para cada modelo entrenado iba en aumento. A su vez, que se analizaba el impacto del número de temas en los modelos LDA, de forma general.

Se puede observar una representación general de la metodología implementada en la SPS en la figura **4-2**. Esta SPS se hizo con el objetivo de encontrar la mejor configuración para las etapas de preprocesamiento, la mejor representación textual y la mejor parametrización del modelo LDA, a través de, un proceso experimental de observación que permitiera inferir el impacto de cada etapa y parámetro según el contexto del conjunto de datos y método LDA.

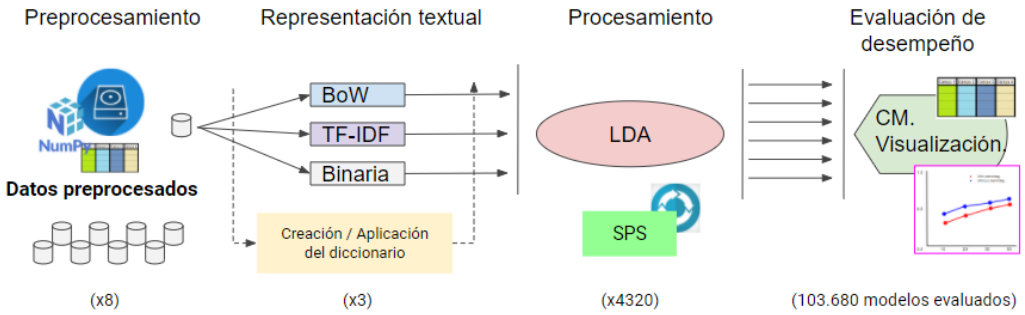


Figura 4-2.: Metodología de SPS general. Elaboración propia.

4.4.2. Diseño de instrumento de evaluación cualitativa

El instrumento de evaluación cualitativa contiene lo siguiente: Un mensaje introductorio de presentación y explicación general del instrumento, aspectos a evaluar con el siguiente orden: *i)* Valoración por inferencia temática (V1), *ii)* Valoración de la asociatividad (V2) y *iii)* Valoración por grado de representatividad (V3), cada evaluación tiene su título, descripción y actividad a realizar; y un mensaje final de agradecimiento con la opción de realizar una nueva evaluación. Se definieron tres aspectos a evaluar por medio del instrumento:

Valoración por inferencia temática (V1)

Los expertos interpretan los términos mostrados para inferir al menos un posible tema, o tantos según considere, que representa mejor dichos términos. Aquí se considera la opinión libre del experto, y según su conocimiento y experiencia pueda relacionar un conjunto de términos para inferir al menos un tema asociado. Un ejemplo para esta valoración es, a partir de los términos "Ratón", "Teclado", "Programa" y "Pantalla" es posible inferir y asociar "Computador" como un tema latente. Para ello se muestran el top-10 de términos más probables para un determinado tema latente obtenido por el modelo LDA. Cada experto evalúa una única vez un tema latente. Respecto al criterio de selección del tema, se prioriza aquellos temas latentes asociados a documentos que pertenecen a la misma área del experto, caso contrario, se muestra uno aleatoriamente. Adicionalmente, cada experto señala el grado de experticia que él considera tiene en los temas propuestos. Para ello, se usó una escala de Likert con los siguientes cuatro niveles "Nada experto", "Poco experto", "Experto" y "Muy experto".

Valoración de la asociatividad (V2)

Se muestra a cada experto un conjunto de términos quienes indican cuáles términos se encuentran más asociados a un artículo científicos mostrado. Cada experto puede asociar desde ninguno hasta todos los términos del conjunto mostrado. Previamente, se muestran el título y resumen del artículo científico con el cual se asociarán los términos según su contenido.

Adicionalmente, cada experto puede visualizar el PDF original del documento. En esta valoración se busca analizar y evaluar el nivel de asociatividad de los términos indicados por los expertos, obtenidos por el modelo LDA, y el artículo científico mostrado. Por ejemplo, para un artículo científico con el título "*Ciencias de la computación*" y con un breve resumen "*La computación ha evolucionado de tal forma que, permite acoplar tecnologías...*", se asocian los términos "Computación", "Estudio" y "Tecnología" y no "Caballo" y "Río".

Adicionalmente, cada experto señala el grado de coherencia que él considera existe entre los términos asociados y el artículo científico. Para ello, se usó una escala de Likert con los siguientes cuatro niveles "Nada coherente", "Poco coherente", "Coherente" y "Muy coherente". En la valoración de la asociatividad se muestra un conjunto de 20 términos, los cuales corresponden a cinco términos de los cuatro temas más probables para el artículo científico mostrado. Por último, cada experto accede únicamente a artículos científicos pertenecientes a su área de estudio y una única vez por artículo. En pro de minimizar el tiempo requerido por cada experto en el diligenciamiento del instrumento, se muestra el mismo artículo científico en las valoraciones de la asociatividad (V2) y por grado de representatividad (V3).

Valoración por grado de representatividad (V3)

Finalmente, se muestra un conjunto de términos de los cuales cada experto selecciona y ordena descendentemente los términos según consideren representan mejor un artículo científico mostrado previamente, donde el primer término será el más representativo y el último, será el menos representativo. Cada experto debe seleccionar entre 1 y 5 términos del conjunto de términos. Previamente, se muestran el título y resumen del documento con el cual se evaluarán los términos según su contenido. Adicionalmente, cada experto puede visualizar el correspondiente PDF original. El objetivo de esta valoración es evaluar el conjunto de términos y orden de los términos seleccionados según su relación con el artículo científico mostrado. El conjunto de términos fue obtenido a partir del modelo LDA. Por ejemplo, para el mismo artículo científico del ejemplo anterior se seleccionan y ordenan los siguientes términos así *i)* "Tecnología", *ii)* "Computador" y *iii)* "Análisis" descartando "Animal" y "Río".

Adicionalmente, cada experto señala el grado de coherencia que él considera existe entre el conjunto de términos (incluyendo términos seleccionados y no seleccionados) y el artículo científico. Para ello, se usó una escala de Likert con los siguientes cuatro niveles "Nada coherente", "Poco coherente", "Coherente" y "Muy coherente". El conjunto de términos mostrado en esta valoración contiene 25 términos, los cuales corresponden a los primeros 25 términos únicos con mayor peso ponderado. El peso ponderado de cada término es igual al producto de la probabilidad del término para un determinado tema por el valor de coherencia

de dicho tema, véase la ecuación 4-3.

$$w_t = p_A(t) * CM_A \tag{4-3}$$

donde w_t es el peso ponderado del término t, $p_A(t)$ es la probabilidad del término t para el tema A, siendo que $t \in$ top-10 de términos probables del tema A y CM_A es el valor de coherencia individual para el tema A.

Por último, cada experto accede únicamente a artículos científicos pertenecientes a su área de estudio y una única vez por artículo. En pro de minimizar el tiempo requerido por cada experto en el diligenciamiento del instrumento, se muestra el mismo artículo científico en las valoraciones de la asociatividad (V2) y por grado de representatividad (V3).

4.4.3. Metodología de aplicación del instrumento

El objetivo de la evaluación cualitativa es analizar y comparar relaciones propuestas por personas con amplio conocimiento y experiencia en determinadas áreas, denominados "expertos". Estas relaciones son comparadas y evaluadas con los resultados de los temas asociados a los documentos obtenidos por el modelo entrenado de LDA. Para ello, se desarrolló una metodología para la aplicación del instrumento de evaluación cualitativa descrito en la sección anterior con los resultados del modelado de temas (TM) con LDA. La metodología inicia desde la búsqueda de información de expertos hasta la aplicación del instrumento de evaluación cualitativa, su consolidación de resultados y análisis. En la figura **4-3** se describe las etapas de la metodología, las cuales se describen a continuación:

1. **Adquirir datos de autores**: Se consolidó en una tabla los datos, correspondientes a los criterios de inclusión, de autores disponibles en las plataformas electrónicas de SCOPUS[3] y Scienti[4] de MinCiencias. Para esta etapa se creó dos grupos de autores, los expertos internos son aquellos que fueron autores o coautores de publicaciones que tienen filiación a la Universidad de los Llanos y cuyos documentos fueron parte del conjunto de datos procesado. Los expertos externos son autores o investigadores que no fueron autores o coautores de publicaciones con filiación a la Universidad de los Llanos y que, por tanto, no fueron parte del conjunto de datos procesado. Los criterios de inclusión fueron existencia de CVLAC, conocimiento del idioma inglés, perfil en SCOPUS, publicaciones recientes no mayores a cinco años a la fecha de realización

[3]Enlace: https://www.scopus.com/search/form.uri?display=basic
[4]Enlace: https://sba.minciencias.gov.co/Buscador_HojasDeVida/

del proceso de recopilación de datos, nivel de estudio posgrado de maestría o superior, contar con categorización como investigador en MinCiencias, tener información de contacto, asignación manual a un área de SCOPUS, poseer trabajos de grado de su formación o producción intelectual relacionados al área de conocimiento. Se tomaron en cuenta únicamente las siete áreas de SCOPUS con mayor cantidad de documentos relacionados (véase la figura **3-6**), estas áreas fueron: "*Agricultural and Biological Sciences*", "*Engineering*", "*Computer Science*", "*Physics and Astronomy*", "*Enviromental Science*", "*Biochemistry, Genetics and Molecular Biology*" y "*Mathematics*". Esto dada la poca cantidad de documentos relacionados para las áreas posteriores.

2. **Filtrar expertos**: Se descartaron los expertos que no cumplieron con todos los criterios de inclusión para ser seleccionados. Para el caso de los expertos externos, la búsqueda se hizo seleccionando expertos que fueran cumpliendo con todos los criterios. Dada la cantidad limitada de expertos internos el filtro se aplicó a todos los existentes, a diferencia de los expertos externos cuyo número máximo de expertos seleccionados se estableció en 20 por área de conocimiento de SCOPUS.

3. **Jerarquizar expertos**: Se jerarquizó los expertos en cada área según el orden de los siguientes criterios: por vinculación directa con la Universidad de los Llanos (perteneciente o no), por categorización de MinCiencias (Investigador Junior, Asociado, Senior y Emérico) y por orden alfabético.

4. **Envío de invitaciones a evaluar**: Se envió una invitación por correo electrónico para participar en la evaluación en grupos de 10 expertos, con un tiempo de espera de respuesta de cinco a diez días.

5. **Envío de enlaces personales**: Para los expertos que respondieron afirmativamente a la invitación se les generó un registro para acceder al instrumento de evaluación cualitativa implementado en la aplicación Web desarrollada "whatTopic" por medio de un enlace único y personal. Esto se hizo para agilizar el proceso de evaluación de cada experto sin recurrir a interfaces de inicio de sesión y para su diligenciamiento completamente independiente y personal.

6. **Diligenciamiento del instrumento**: Se estableció un mínimo de evaluaciones por área, siendo mínimo dos evaluaciones de expertos internos y una de externos. Cada evaluación se debe resolver una única vez por experto en determinada área en la cual se encuentra clasificado.

7. **Evaluación de resultados**: Por medio de métodos y relaciones, como el análisis de los datos por medio del método de codificación, representaciones y agrupaciones visuales como nubes de términos, tablas y análisis estadístico, se evaluaron y analizaron los resultados cualitativos de la evaluación por medio del instrumento con los expertos. El

método de codificación, en un análisis cualitativo, consiste en segmentar o fragmentar los datos (términos propuestos por los expertos en este caso) por medio de códigos o categorías en función de su significado interpretativo acorde al objeto de análisis e investigación. Estos códigos son en esencia los datos de forma condensada en unidades analizables que permiten relacionar conceptos comunes o ideas [González Gil and Cano Arana, 2010, Acevedo, 2011].

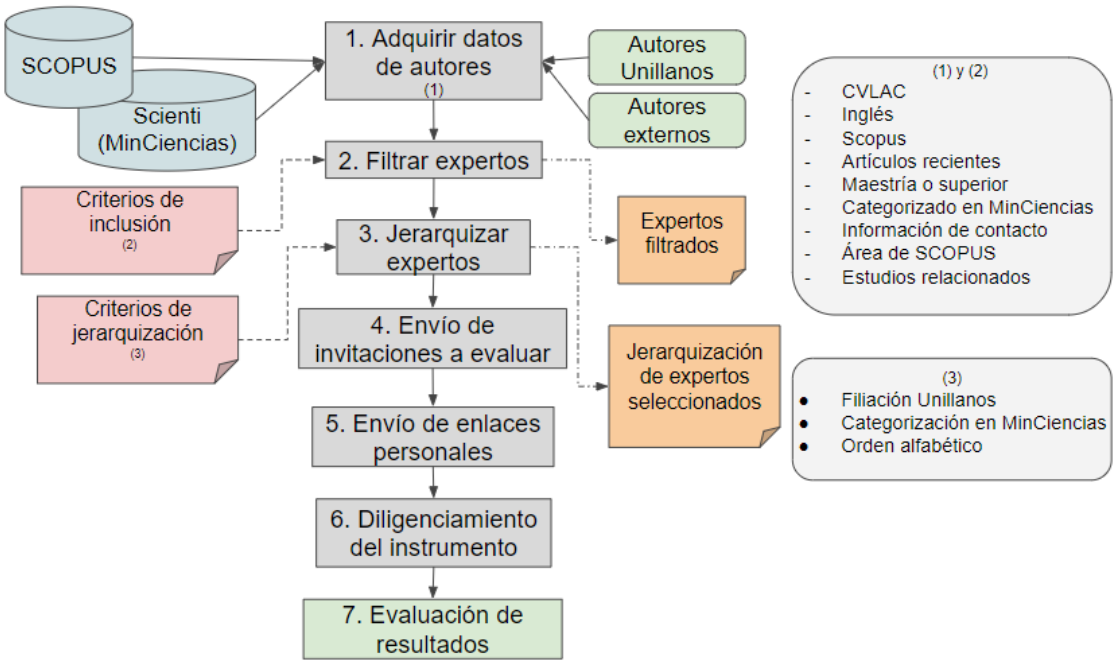


Figura 4-3.: Metodología de aplicación del instrumento de evaluación cualitativa. Elaboración propia.

5. Evaluación y resultados

En este capítulo se presentan, describen y evalúan los resultados obtenidos por el proceso de evaluación cuantitativa y, posteriormente, los resultados obtenidos por la evaluación cualitativa.

5.1. Evaluación cuantitativa

5.1.1. Resultados

Búsqueda Sistemática de Parámetros (SPS)

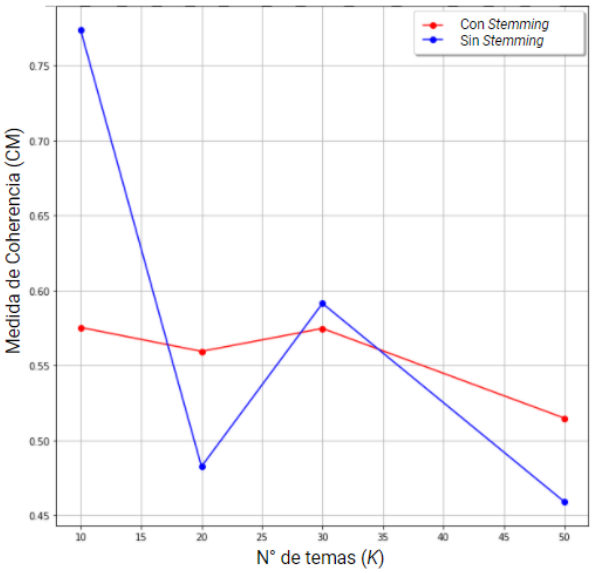
La tabla **5-1** muestra las mejores configuraciones obtenidas según cada tamaño diferente de ventana deslizante de CM. La figura **5-1** muestra los valores CM para cada uno de los parámetros y etapas a comparar, donde cada valor corresponde a una única ejecución por cada configuración del modelo. Las subfiguras **5-1**a, **5-1**b, **5-1**c y **5-1**d muestran las etapas "*Stemming*", "Creación de Bigramas y Trigramas", "Poda" y la fase "Representación textual", respectivamente. Por último, la figura **5-2** muestra la fase de procesamiento y parametrización del modelo LDA.

Al analizar dichos resultados se observó que: *i)* Las configuraciones y aplicaciones de las etapas para el conjunto de datos preprocesado que obtuvieron mayor CM eran diferentes. *ii)* La representación textual TF-IDF presentó, en términos generales, mayores valores CM que las otras dos representaciones textuales, aunque con BoW el aumento no era considerable. *iii)* Los parámetros de los modelos LDA variaban en los resultados obtenidos. *iv)* No había una relación entre mejores resultados y configuración de tamaño de ventana deslizante. Por tal motivo, se concluyó que:

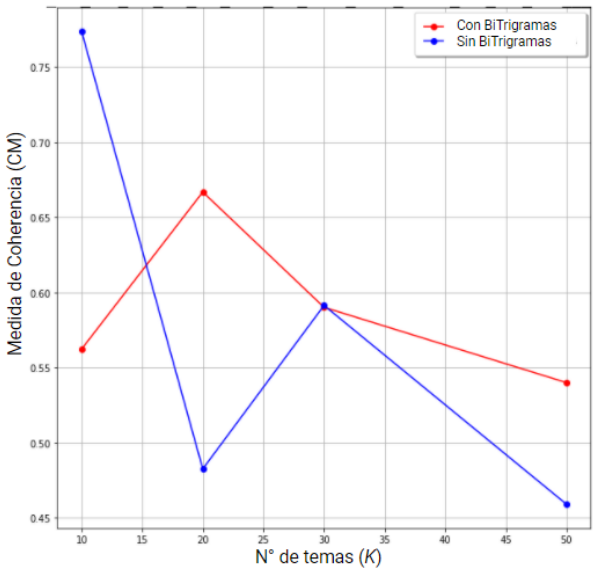
1. Aún no es posible dar un criterio concluyente y claro para escoger la mejor configuración de preprocesamiento.

2. A pesar de obtener la representación TF-IDF para cada configuración de LDA, se hace necesario realizar nuevamente comparaciones con BoW y Matriz Binaria.

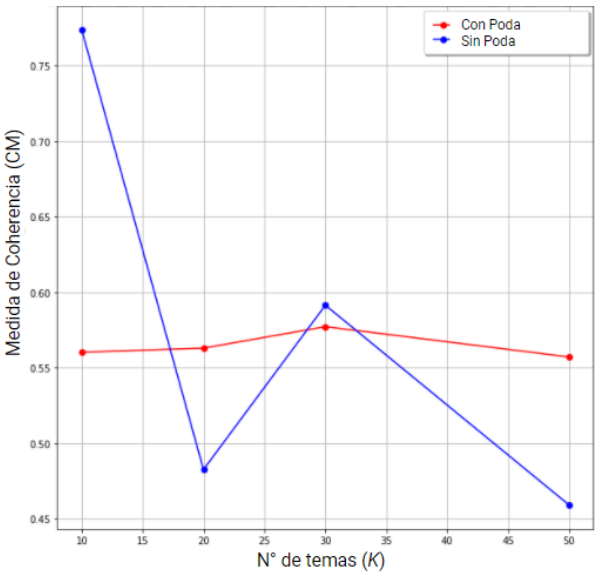
3. Los parámetros más relevantes de LDA serían los directamente definidos en el método LDA (α, η y K), mientras que el resto no se consideró de mayor impacto por lo que



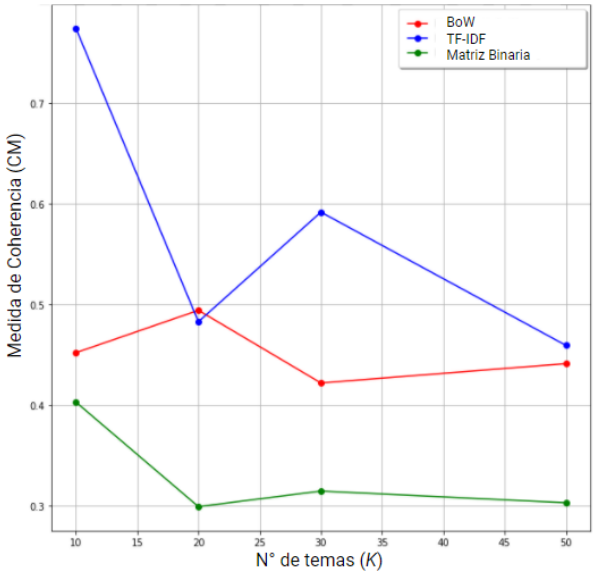
(a) Etapa "*Stemming*".



(b) Etapa "Creación de Bigramas y Trigramas".



(c) Etapa "Poda".



(d) "Representación textual" (BoW, TF-IDF y Matriz binaria).

Figura 5-1.: Comparación con y sin cada etapa de preprocesamiento y la representación textual variando el número de temas. Elaboración propia.

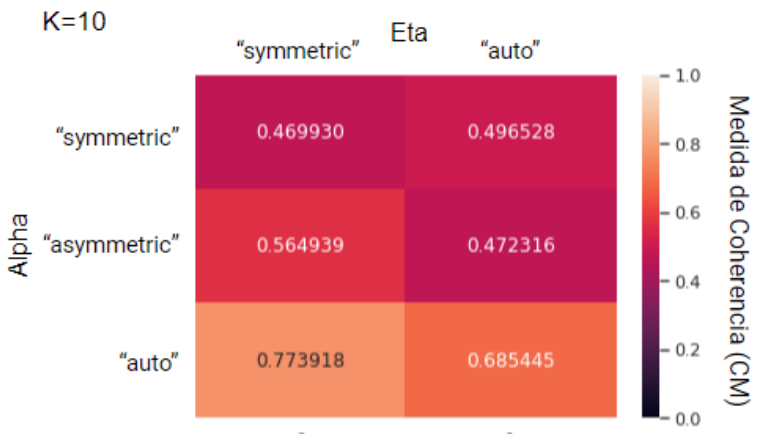


Figura 5-2.: Comparación hiperparámetros ***alpha*** y ***eta*** para un modelo LDA con 10 temas ($K = 10$). Elaboración propia.

se toma la configuración de los demás parámetros de acuerdo al cuarto modelo, con la CM más alta ($CM = 0.773$). Estos parámetros quedarían configurados de la siguiente manera: "***chunksize***":10, "***passes***":15, **LDA**: Batch, "***decay***":1, "***iterations***":25.

4. El tamaño de la ventana deslizante no ocasionaba diferencias relevantes entre resultados. Por lo que se implementó la configuración estándar establecida por [Röder et al., 2015] para un tamaño de 110.

Tabla 5-1.: Mejores resultados de la SPS en relación a la medida de Coherencia por cada tamaño de ventana deslizante

Configuración/Parametrización			**Tamaño ventana CM**			
Conjunto de datos	**Repre. textual**	**Modelo LDA**	**10**	**25**	**50**	**110**
"***stemming***": Sí, **BiTrigramas**: Sí, **Poda**: Sí.	TF-IDF	"***alpha***":"auto", "***eta***":"auto", **No. de temas** (K):50, "***chunksize***":10, "***passes***":7, **LDA**:Batch, "***decay***":0.51, "***iterations***":25	**0.719**	0.687	0.623	0.570
"***stemming***": Sí, **BiTrigramas**: Sí, **Poda**: Sí.	TF-IDF	"***alpha***":"auto", "***eta***":"auto", **No. de temas** (K):50, "***chunksize***":10, "***passes***":7, **LDA**:Batch, "***decay***":0.51, "***iterations***":25	0.719	**0.687**	0.623	0.570
"***stemming***": No, **BiTrigramas**: No, **Poda**: Sí.	TF-IDF	"***alpha***":"asymmetric", "***eta***":"auto", **No. de temas** (K):10, "***chunksize***":137, "***passes***":1, **LDA**:Batch, "***decay***":0.51, "***iterations***":50	0.582	0.661	**0.710**	0.733
"***stemming***": No, **BiTrigramas**: No, **Poda**: No.	TF-IDF	"***alpha***":"auto", "***eta***":"symmetric", **No. de temas** (K):10, "***chunksize***":10, "***passes***":15, **LDA**:Batch, "***decay***":1, "***iterations***":25	0.503	0.589	0.675	**0.773**

Por tal razón, se realizó una segunda SPS donde por cada combinación de los parámetros pendientes se generaba cinco modelos LDA entrenados para ser evaluados y cuyos valores

CM se graficaron para analizar cada configuración y su impacto. Las figuras desde **5-3** hasta **5-8** muestran el proceso de comparación tanto por etapa de preprocesamiento como parámetros LDA. Por ejemplo, la etapa "***Stemming***" se puede observar y comparar tanto en la figura **5-3** como en la figura **5-4**. Adicionalmente, en dichas figuras se comparan los parámetros ***alpha*** y ***eta*** respectivamente, para esta etapa. Esto permitió analizar el impacto tanto de las etapas de preprocesamiento como de los parámetros LDA para cada una de las representaciones textuales. En la tabla **5-2** se muestra los valores de los parámetros destinados a esta segunda SPS. Cada una de las figuras de la **5-3** a **5-8** tienen tres columnas, donde cada una hace referencia a un tipo de representación textual (BoW, TF-IDF, Binaria). Por otra parte, cada fila corresponde a las opciones de los valores del parámetro a evaluar y comparar. Al igual que en la primera SPS, el proceso de comparación consistió en usar la configuración de parámetros base seleccionada y variar únicamente la etapa o parámetro a evaluar. Igualmente, se evaluó el número de temas como variable independiente en el eje x de las gráficas de cada figura, mientras que el eje y correspondió al valor de CM. Cada una de las figuras muestran los valores medios de CM para cada configuración obtenidos de los cinco modelos con el mismo valor del parámetro, representada por una linea gruesa, y su correspondiente desviación estándar, la cual se representa como una sombra alrededor de la linea de la media. Las gráficas de cada una de las figuras de esta SPS están organizadas y presentadas de la siguiente manera:

- Etapa "***Stemming***": La figura **5-3** varía el parámetro "**alpha**" entre los valores ("symmetric", "asymmetric" y "auto") y la figura **5-4** el parámetro "**eta**" con los siguientes valores ("symmetric" y "auto").
- Etapa "**Creación de Bigramas y Trigramas**": La figura **5-5** varía el parámetro "**alpha**" entre los valores ("symmetric", "asymmetric" y "auto") y la figura **5-6** el parámetro "**eta**" con los siguientes valores ("symmetric" y "auto").
- Etapa "**Poda**": La figura **5-7** varía el parámetro "**alpha**" entre los valores ("symmetric", "asymmetric" y "auto") y la figura **5-8** el parámetro "**eta**" con los siguientes valores ("symmetric" y "auto").

En la figura **5-3** se puede observar que en la mayoría de los casos y por poca diferencia, se desempeña mejor los modelos con *stemming*. Además, las desviaciones estándar en todos los casos eran similares en modelos con o sin *stemming*. Respecto a la configuración para "**alpha**", el valor "auto" presentó mejores resultados. Por otra lado, las representaciones BoW y TF-IDF obtuvieron mejores desempeños notables en comparación con Matriz Binaria. Entre estas dos representaciones presentaron desempeños similares exceptuando en la configuración con "**alpha**" con valor "auto" el cual TF-IDF tuvo mejor desempeño, no obstante, su desviación estándar era igualmente mayor. Por último, se observó que al aumentar el número de temas disminuía o no aumentaba sustancialmente el desempeño.

Tabla 5-2.: Valores de los parámetros para una segunda SPS por cada etapa

Preprocesamiento	
Etapa	**Aplica**
Stemming	Sí/No
Creación de Bigramas y Trigramas	Sí/No
Poda	Sí/No
Representación textual	
Tipo	[BoW; TF-IDF; Binaria]
Procesamiento	
Parámetro	**Valor**
"**alpha**"	["symmetric"; "asymmetric"; "auto"]
"**eta**"	["symmetric"; "auto"]
"**num_topics**"	[10; 20; 30; 50]

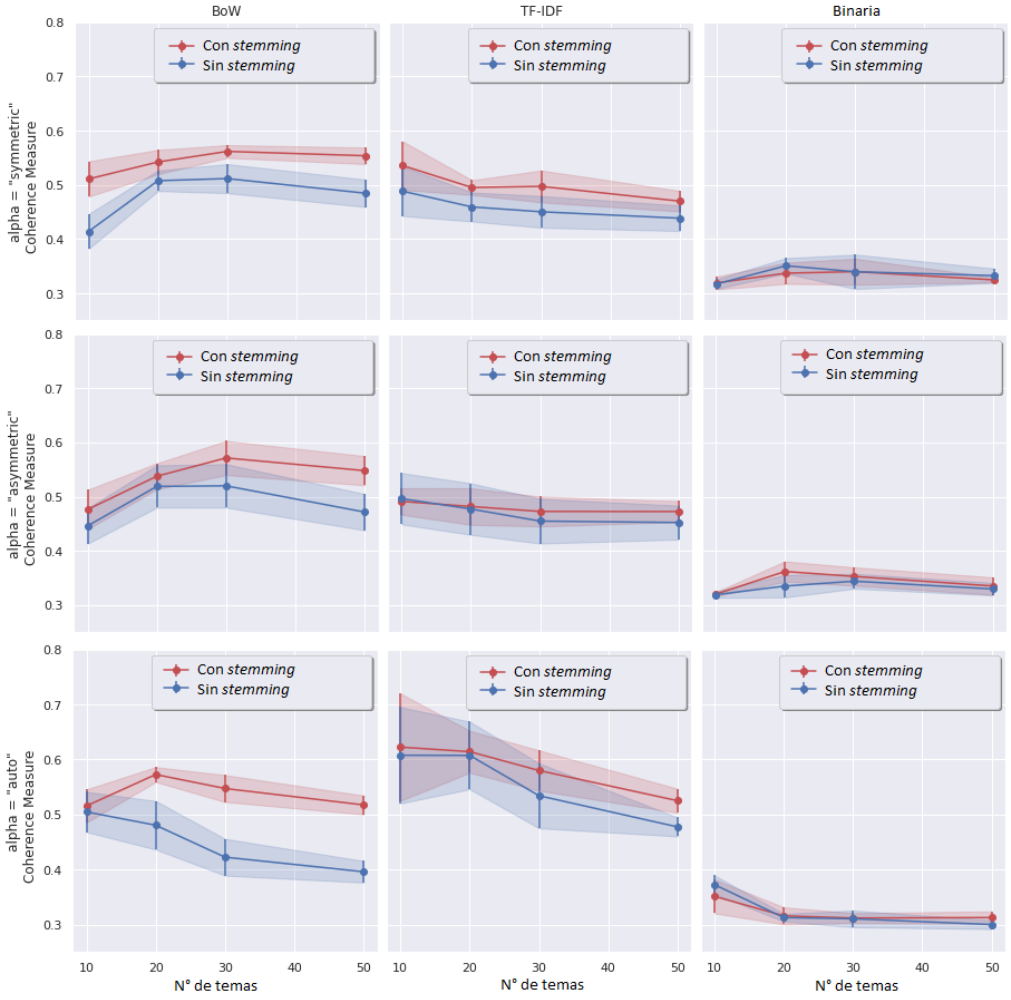


Figura 5-3.: Comparación con y sin "*stemming*" variando el parámetro "alpha" ("symmetric", "asymmetric" y "auto"), para cada una de las representaciones textuales (BoW, TF-IDF, Binaria). Elaboración propia.

Seguidamente, en la figura **5-4** se observa que los modelos con *stemming* tuvieron un mejor desempeño, con un aumento no sustancial, con aquellos en los que no se aplicó. También, se observó que en un número bajo de temas (i.e. 10) ambos tipos de modelos presentaron desempeños similares, a excepción de un caso con "**eta**" con valor "auto", donde los modelos sin *stemming* demostraron tener mejor desempeño. Al igual que en la figura anterior, La Matriz Binaria obtuvo los desempeños más bajos, la TF-IDF obtuvo mejor desempeño que BoW pero sin ser considerable. La desviación estándar en TF-IDF sigue siendo mucho mayor, principalmente en un número de temas menor (i.e. 10, 20) comparada con el resto de representaciones textuales.

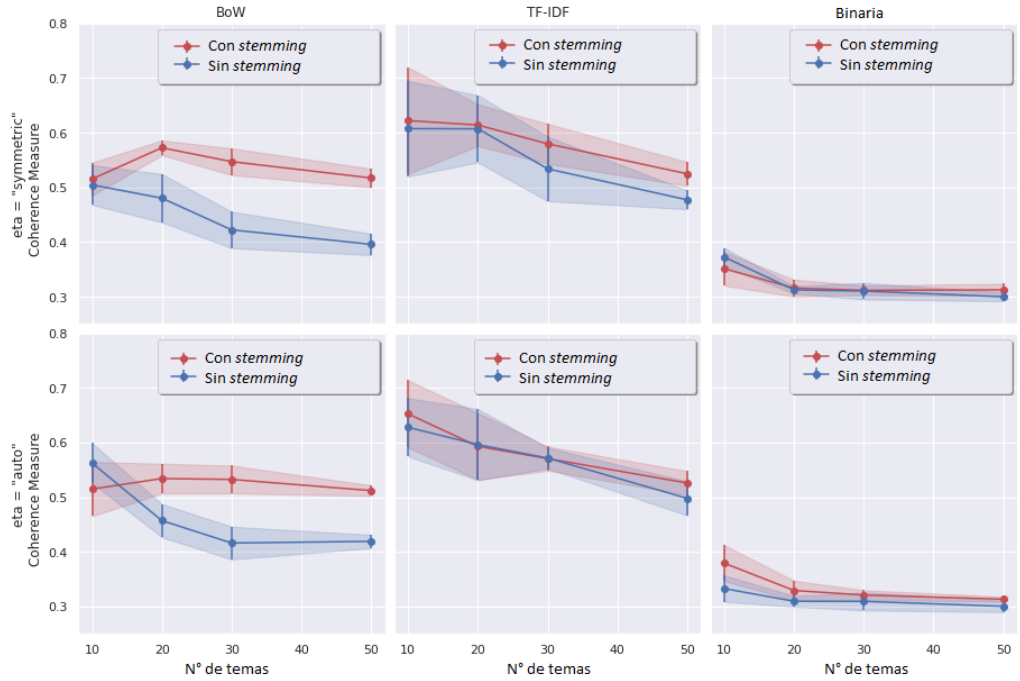


Figura 5-4.: Comparación con y sin "*stemming*" variando el parámetro "eta" ("symmetric" y "auto"), para cada una de las representaciones textuales (BoW, TF-IDF, Binaria). Elaboración propia.

A continuación, en la figura **5-5** se muestra que los modelos con y sin "Creación de Bigramas y Trigramas" obtuvieron desempeños similares. Para BoW los modelos sin la aplicación de esta etapa obtuvieron levemente mejores desempeños. Por otro lado, en TF-IDF fueron los modelos con esta aplicación quienes se desempeñaron mejor con un aumento mínimo. En términos generales, la implementación o no de esta etapa resultó con desempeños similares. Respecto a la configuración del parámetro "**alpha**", el valor "auto" se desempeñó mejor a pesar que presentar una mayor desviación estándar en comparación con los otros valores. Por último, la representación textual que obtuvo mayores desempeños fue TF-IDF al igual que obtuvo mayores desviaciones estándar, BoW presentó menores desviaciones sin disminuir

tanto el desempeño.

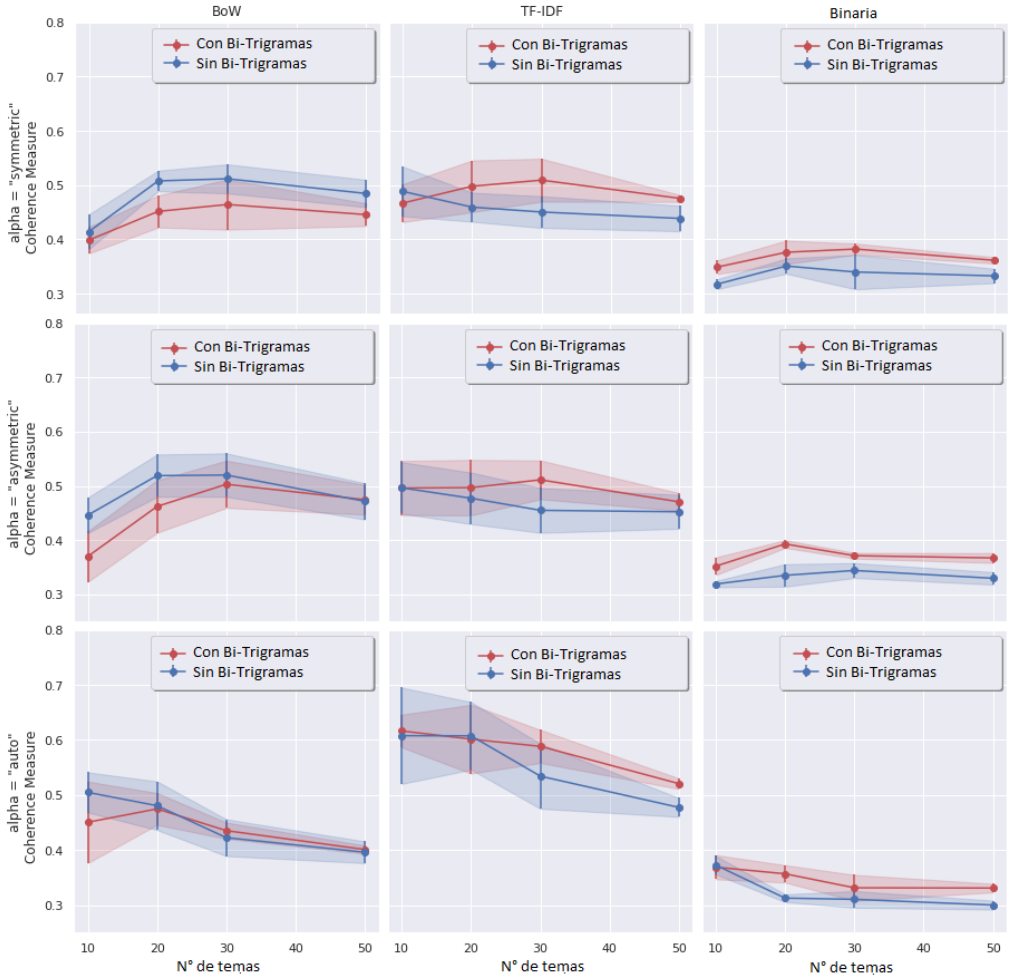


Figura 5-5.: Comparación con y sin "Creación de Bigramas y Trigramas" variando parámetro "alpha" ("symmetric", "asymmetric" y "auto"), para cada una de las representaciones textuales (BoW, TF-IDF, Binaria). Elaboración propia.

Posteriormente, en la figura **5-6**, se observa un comportamiento similar a los casos ya descritos, donde la representación textual que mejor desempeño tuvo fue TF-IDF pero cuya desviación estándar igualmente fue mayor. La representación BoW presentó un desempeño menor que TF-IDF pero con menos desviación. Los modelos con la aplicación de esta etapa mostraron un leve mejor desempeño en TF-IDF y Matriz Binaria, para el caso de BoW, tuvieron mejor desempeño los modelos que no aplicaron esta etapa. Por último, los valores asignados a "**eta**" obtuvieron desempeños sin diferencia destacable.

Luego, en la figura **5-7** se puede observar que los valores de "**alpha**" obtuvieron desempeños similares, a excepción de "auto" con la representación TF-IDF la cual tuvo mayores

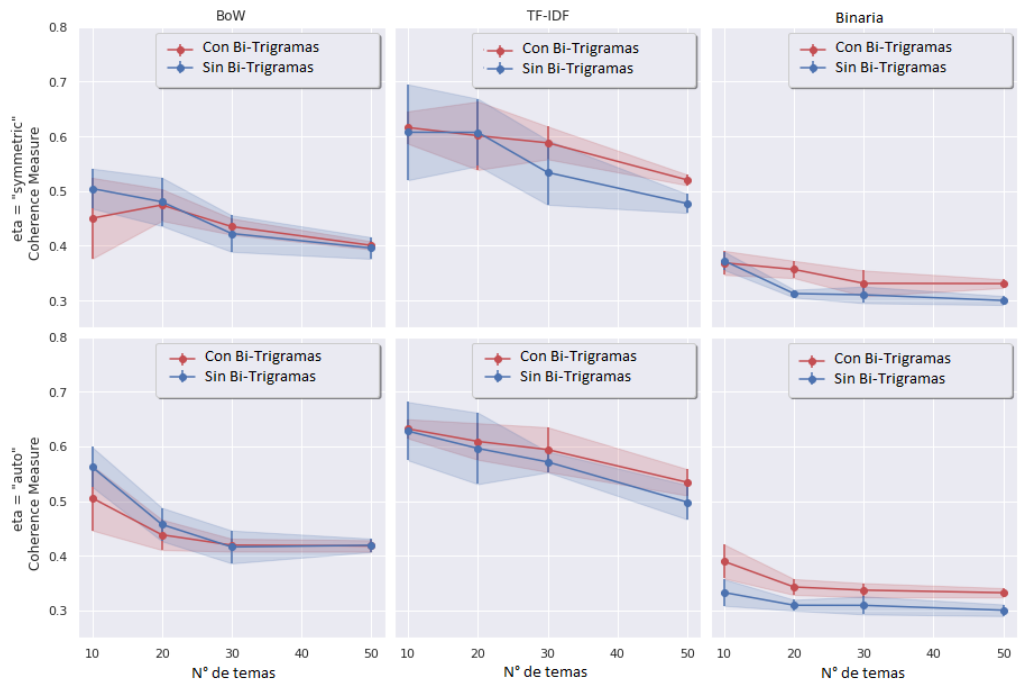


Figura 5-6.: Comparación con y sin "Creación de Bigramas y Trigramas" variando parámetro "eta" ("symmetric" y "auto"), para cada una de las representaciones textuales (BoW, TF-IDF, Binaria). Elaboración propia.

desviaciones estándar. Hubo una leve diferencia favorable en los modelos con "Poda" en comparación con aquellos sin "Poda". Reiteradamente, la TF-IDF obtuvo mejores desempeños con mayores desviaciones estándar. La representación BoW tuvo desempeños similares a TF-IDF con menores desviaciones.

Por último, en la figura **5-8** el comportamiento es similar a la figura anterior, donde TF-IDF obtuvo mejores desempeños con mayores desviaciones estándar. BoW seguidamente, presentó menores desviaciones. No se identificó una diferencia notable en los modelos con y sin "Poda". Respecto al valor para "**eta**", presentaron desempeños similares.

Tomando en cuenta las observaciones descritas para cada una de las figuras mencionadas anteriormente, se concluye que la configuración mostrada en la tabla **5-3** es la óptima para el enfoque de este trabajo. En dicha tabla se argumenta la razón por la cual se tomó cada configuración. A partir de esta configuración se obtuvo un modelo LDA que alcanza un valor de desempeño de CM = 0.639. Cabe aclarar que, a pesar de haber obtenido un modelo con un valor de CM mayor en la primera SPS (CM = 0.773) dicho modelo no fue seleccionado, debido a que se tomó en cuenta cada uno de los argumentos descritos en la tabla **5-3**, los cuales relacionaban una configuración de parámetros diferente de la implementada en dicho modelo.

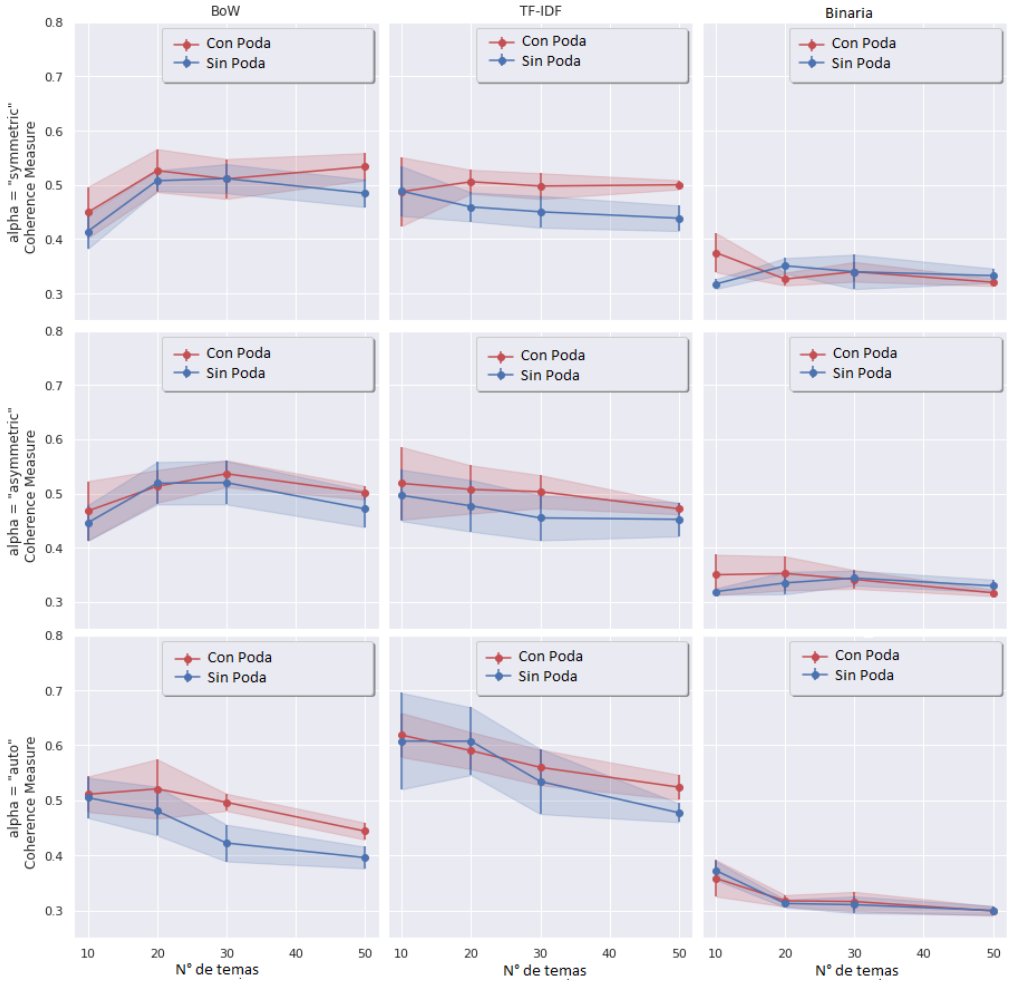


Figura 5-7.: Comparación con y sin "Poda" variando parámetro "alpha" ("symmetric", "asymmetric" y "auto"), para cada una de las representaciones textuales (BoW, TF-IDF, Binaria). Elaboración propia.

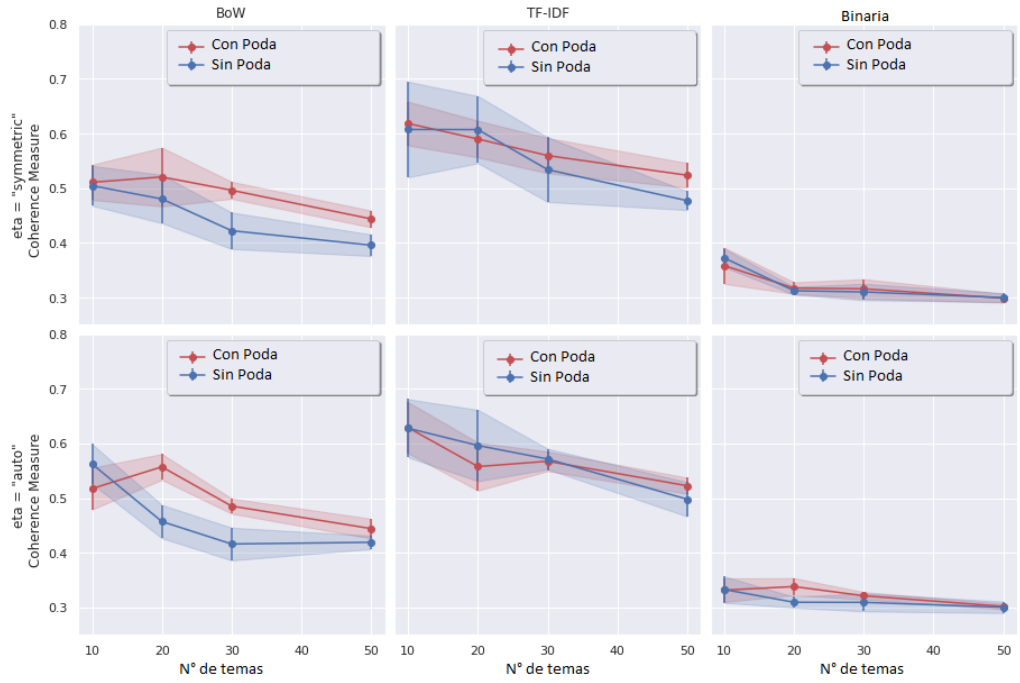


Figura 5-8.: Comparación con y sin "Poda" variando parámetro "eta" ("symmetric" y "auto"), para cada una de las representaciones textuales (BoW, TF-IDF, Binaria). Elaboración propia.

Tabla 5-3.: Resultados segunda SPS. Parametrización/Configuración seleccionada.

Preprocesamiento		
Etapa	**Aplicación**	**Justificación**
Stemming	No	Menor dispersión en modelos con pocos temas.
Creación de Bigramas y Trigramas	Sí	Variación con poco impacto. Se propone para aumentar el número de términos dada la poca cantidad de artículos.
Poda	No	Menor dispersión en modelos con pocos temas.
Representación textual	BoW	Mayor balance entre un buen desempeño y una menor dispersión.
Procesamiento		
Parámetro	**Valor**	**Justificación**
"**alpha**"	"auto"	Mejores desempeños en modelos con pocos temas.
"**eta**"	"symmetric"	Desempeño levemente mejor.
"**num_topics**"	10	Dada la cantidad de artículos científicos (137) y el desempeño observado. El desempeño tendía a disminuir con el aumento de número de temas.

Temas latentes - LDA

El conjunto de datos preprocesado, configuración mostrada en la tabla **5-3**, presenta las siguientes características: Número de artículos científicos: 137. Tamaño del diccionario: 41,719 términos únicos. Un *corpus* con 334,115 términos en total. El modelo LDA obtenido alcanza un valor de desempeño global de $CM_{global} = 0.639$. Consecuentemente, en este trabajo a este valor global se denominó como CM_{global} o simplemente CM. Este modelo se aplicó con la siguiente parametrización: "**alpha**" = "auto", "**eta**" = "symmetric" y número de temas ($\boldsymbol{K}$) = 10. Los 10 temas latentes obtenidos con el modelo LDA son mostrados en la tabla **5-4**, la cual permite observar el top 10 de términos según su probabilidad por tema ordenados descendentemente y el valor individual de coherencia de cada tema CM_i, donde i es el ID del tema. Cada término es mostrado en conjunto con su traducción en Español, según el caso. La barra de color que ilustra la escala representada con el valor de probabilidad de cada término asociado a los colores de la tabla **5-4** se muestra en la figura **5-9**. Para referenciar mejor los temas en las figuras posteriores se propuso una paleta de colores donde cada tema latente tiene asignado un color determinado de dicha paleta. La paleta se puede ver implementada en la tabla **5-4** en las celdas correspondiente a los identificadores de cada tema latente. En la figura **5-10** se muestran los valores CM para cada uno de los 10 temas latentes obtenidos por el modelo LDA. Estos valores presentan un valor medio de 0.639 (siendo el mismo valor CM para el modelo LDA en general), una desviación estándar de 0.1907, un valor máximo de 0.894 para el tema 3 y un mínimo de 0.375 en el tema 2.

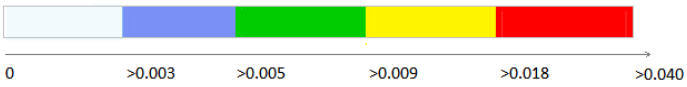


Figura 5-9.: Barra de color que ilustra la escala representada con el valor de probabilidad de cada término asociado a los colores de la tabla **5-4**. Elaboración propia.

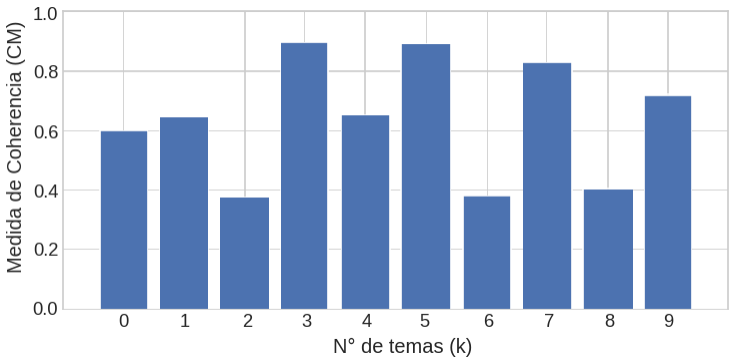


Figura 5-10.: Valores de la medida de coherencia (CM) por cada uno de los 10 temas latentes. Elaboración propia.

Tomando en cuenta los valores de las probabilidades de los términos como el grado de

Tabla 5-4.: Resultados temas latentes, $CM_{global} = 0.639$. Top 10 de los términos más probables para cada tema.

Tema 0 (CM_0 =**0.600**)		**Tema 1** (CM_1 =**0.647**)		**Tema 2** (CM_2 =**0.375**)		**Tema 3** (CM_3 =**0.894**)		**Tema 4** (CM_4 =**0.652**)	
0.006	case (“caso”)	0.017	fish (“pez”)	0.011	starch (“almidón”)	0.006	produce_water (“producir_agua”)	0.009	cattle (“ganado/ganadería”)
0.006	value valor/valorar	0.012	sh	0.006	cellulose (“celulosa”)	0.005	crude_oil (“petróleo_crudo”)	0.007	ipcc
0.005	orbit (“órbita/orbitar”)	0.010	observe (“observar”)	0.004	peak (“pico”)	0.004	algae (“algas”)	0.007	livestock (“ganado/ganadería”)
0.004	order (“orden/ordenar”)	0.009	exposure (“exposición”)	0.003	increase (“aumento/aumentar”)	0.004	chlorophyll (“clorofila”)	0.007	soil (“suelo”)
0.004	correspond (“corresponder”)	0.008	water (“agua”)	0.003	sample (“muestra”)	0.003	treatments (“tratamientos”)	0.005	emissions (“emisiones”)
0.004	present (“presente/presentar”)	0.007	effect (“efecto”)	0.003	properties (“propiedades”)	0.003	cell_density (“densidad_celular”)	0.005	farm (“granja”)
0.004	point (“punto/señalar”)	0.007	concentration (“concentración”)	0.003	compound (“compuesto”)	0.003	pw	0.004	systems (“sistemas”)
0.003	energy (“energía”)	0.005	endosulfan (“endosulfán”)	0.003	cassava_starch (“almidón_de_yuca”)	0.002	pigment (“pigmento”)	0.004	pasture (“pasto/pastar”)
0.003	time (“tiempo”)	0.005	gill (“agallas”)	0.003	nitrogen (“nitrógeno”)	0.002	chlorella_vulgaris	0.004	variables (“variables”)
0.003	potential (“potencial”)	0.005	liver (“hígado”)	0.003	higher (“mayor/más_alto”)	0.002	growth (“crecimiento”)	0.004	emission_factor (“factor_de_emisión”)
Tema 5 (CM_5 =**0.890**)		**Tema 6** (CM_6 =**0.379**)		**Tema 7** (CM_7 =**0.826**)		**Tema 8** (CM_8 =**0.404**)		**Tema 9** (CM_9 =**0.722**)	
0.002	mq	0.004	result (“resultado”)	0.006	tomato (“tomate”)	0.013	species (“especies”)	0.020	soil (“suelo”)
0.001	pid_controller (“controlador_pid”)	0.004	study (“estudio/estudiar”)	0.004	eccentricity (“excentricidad”)	0.008	colombia	0.009	metal (“metal”)
0.001	bay_jamaica (“bahía_jamaica”)	0.003	sample (“muestra”)	0.003	fruit (“fruta”)	0.007	en	0.008	heavy_metal (“metal_pesado”)
0.001	small_shed (“pequeño_cobertizo”)	0.003	different (“diferente”)	0.003	contour (“contorno”)	0.003	study (“estudio/estudiar”)	0.005	sediment (“sedimento”)
0.001	cities_communities _july_montego (“ciudades_comunidades _julio_montego”)	0.003	increase (“aumento/aumentar”)	0.002	object (“objeto”)	0.003	del	0.005	urban (“urbano”)
0.001	industry_innovation _infrastructure_sustainable (“industria_innovación _infrastructura_sustentable”)	0.003	model (“modelo/modelar”)	0.002	actual (“actual”)	0.003	specimens (“especímenes”)	0.005	mg
0.001	gas (“gas”)	0.003	value (“valor/valorar”)	0.002	tomato_fruit (“tomate_fruta”)	0.003	bird (“ave”)	0.004	water (“agua”)
0.001	arduino	0.003	effect (“efecto”)	0.002	pixels (“pixeles”)	0.003	record (“registro/registrar”)	0.004	source (“fuente”)
0.001	poultry (“aves_de_corral”)	0.003	use (“uso/usar”)	0.002	tomato_size (“tamaño_de_tomate”)	0.003	include (“incluye/incluir”)	0.004	adsorption (“adsorción”)
0.001	sns_mq	0.003	data (“dato”)	0.002	statistical_moments (“momentos_estadísticos”)	0.002	base (“base”)	0.004	wastewater (“aguas_residuales”)

pertenencia de los términos al tema, de acuerdo con la tabla **5-4**, en la figura **5-11** se visualiza una nube de términos por tema donde los términos de mayor tamaño son considerados los de mayor relación con su respectivo tema.

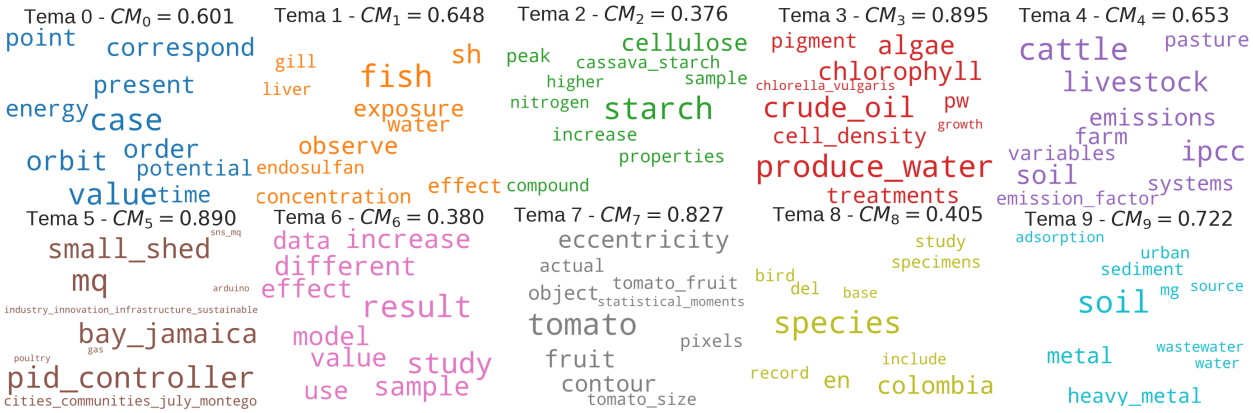


Figura 5-11.: Nube de términos por tema de acuerdo con su contribución. Elaboración propia.

Distribución de documentos

En la tabla **5-5** se puede observar un fragmento de los tres primeros documentos de ejemplo, cada uno asociado a un tema, de acuerdo con la probabilidad más alta o grado de pertenencia del documento a un tema. Esta tabla muestra el identificador (ID) del documento, el identificador del tema dominante, contribución del tema en el documento, los 10 términos con mayor probabilidad del tema dominante, el título del documento y el top 10 de términos del documento según el tema dominante. La tabla completa se puede consultar en el anexo B.

Tabla 5-5.: Tema dominante por documento. Fragmento.

ID Doc.	Tema dominante	Prob. Tema [1]	Términos Top-10 del tema	Título del documento	Términos Top-10 del Doc. según el Tema Dominante
0	6	0.7720	result, study, sample, different, increase, model, value, effect, use, data	"*Monitoring system of relativy humidity, CO, CO2, NH3 and temperature control for small shed*"	[control, design, response, phase, temperature, sensor, source_author, level, environment, necessary]
1	6	0.7290	result, study, sample, different, increase, model, value, effect, use, data	"*Algorithms to estimation of size and shape tomato using Artificial Vision Techniques*"	[image, obtain, size, implement, application, process, calculate, algorithm, classification, define]
2	0	0.9960	case, value, orbit, order, correspond, present, point, energy, time, potential	"*Numerical and analytical analysis of a 3UPS-2RPRRR parallel robot*"	[point, analysis, solution, value, base, equations, platform, cos, vector, sin]

[1] Probabilidad del tema en el documento

En la tabla **5-6** se muestra el documento más representativo, con la probabilidad más alta, para cada uno de los 10 temas latentes.

Para identificar la distribución de los términos asociados a los temas en cada documento, así como el tema dominante por documento, se presenta la figura **5-12**, la cual contiene una muestra de 10 documentos donde se representa el tema dominante de cada documento con un cuadrado con el color asociado al tema y los primeros términos coloreados según el tema asociado a cada uno.

En las figuras **5-13**, **5-14** y **5-15** se presentan en detalle fragmentos de tres documentos de ejemplo (Doc. 2, Doc 7. y Doc. 10) relacionados a tres temas dominantes diferentes (tema 0, tema 6 y tema 2) respectivamente, donde se puede identificar como la mayoría de términos están justamente asociados al tema con mayor probabilidad asociada al documento.

Visualización de temas

La visualización de temas usando la representación bidimensional de los temas basado en MDS con pyLDAvis se presenta en la figura **5-16**. Cada tema es presentado con su identificador y algunos de sus primeros términos de mayor probabilidad asociada que lo conforman.

Adicionalmente, se aplicó el método t-SNE por medio del módulo Scikit-learn [Pedregosa et al., 2011], para visualizar los documentos en un espacio bidimensional y señalar los temas

Tabla 5-6.: Documento representativo por tema.

ID Tema	Prob. Tema[1]	Términos Top-10 del tema	ID Doc. Rep.[2]	Título del documento
0	0.996	case, value, orbit, order, correspond, present, point, energy, time, potential	2	"*Numerical and analytical analysis of a 3UPS-2RPRRR parallel robot*"
1	0.661	fish, sh, observe, exposure, water, effect, concentration, endosulfan, gill, liver	3	"*Biochemical and histological alterations in Aequidens metae (Pisces, Cichlidae) and Astyanax gr. bimaculatus (Pisces, Characidae) as indicators of river pollution*"
2	0.891	starch, cellulose, peak, increase, sample, properties, compound, cassava_starch, nitrogen, higher	9	"*Harnessing CO 2 into Carbonates Using Heterogeneous Waste Derivative Cellulose-Based Poly(ionic liquids) as Catalysts*"
3	0.384	produce_water, crude_oil, algae, chlorophyll, treatments, cell_density, pw, pigment, chlorella_vulgaris, growth	7	"*Physiological and enzymatic responses of Chlorella vulgaris exposed to produced water and its potential for bioremediation*"
4	0.703	cattle, ipcc, livestock, soil, emissions, farm, systems, pasture, variables, emission_factor	6	"*Emission factors estimated from enteric methane of dairy cattle in Andean zone using the IPCC Tier-2 methodology*"
5	0.201	mq, pid_controller, bay_jamaica, small_shed, cities_communities_july_montego, industry_innovation_infrastructure_sustainable, gas, arduino, poultry, sns_mq	0	"*Monitoring system of relativy humidity, CO, CO2, NH3 and temperature control for small shed*"
6	0.991	result, study, sample, different, increase, model, value, effect, use, data	72	"*A Bayesian inference method for estimating the channel occupancy*"
7	0.208	tomato, eccentricity, fruit, contour, object, actual, tomato_fruit, pixels, tomato_size, statistical_moments	1	"*Algorithms to estimation of size and shape tomato using Artificial Vision Techniques*"
8	0.985	species, colombia, en, study, del, specimens, bird, record, include, base	25	"*A New Species of Spatuloricaria Schultz, 1944 (Siluriformes: Loricariidae), from the Orinoco River Basin, Colombia*"
9	0.763	soil, metal, heavy_metal, sediment, urban, mg, water, source, adsorption, wastewater	8	"*Land-use-dependent spatial variation and exposure risk of heavy metals in road-deposited sediment in Villavicencio, Colombia*"

[1] Probabilidad del tema en el documento

[2] ID del documento representativo

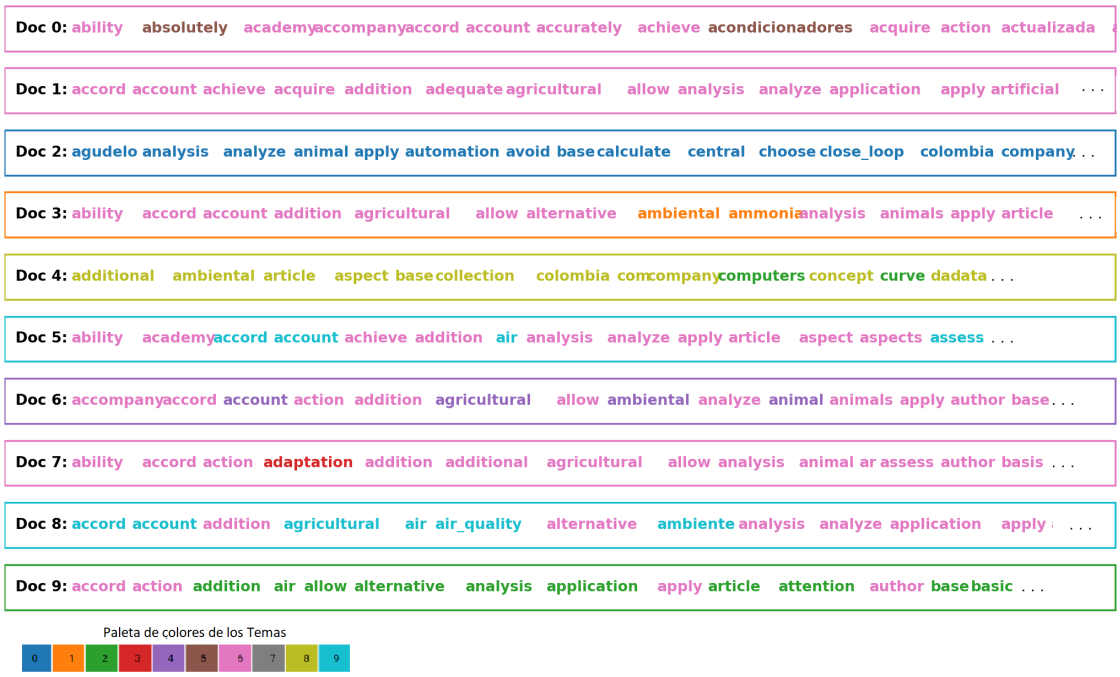


Figura 5-12.: Muestra de documentos según su tema dominante representado por el borde del cuadrado y sus respectivos términos asociados de acuerdo a relación con cada tema en el documento. Elaboración propia.

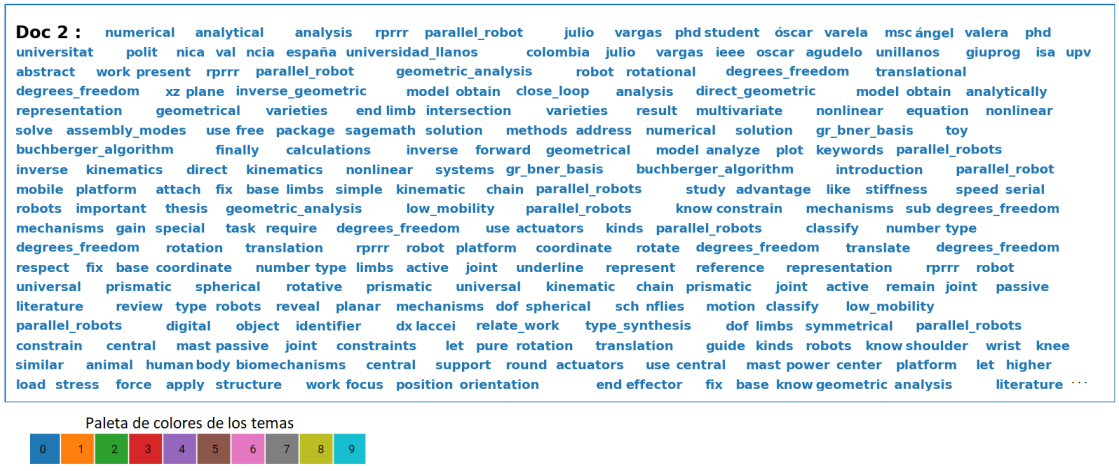


Figura 5-13.: Fragmento del documento 2 según con el tema 0 como dominante representado por el color azul, así como el tema codificado por color de cada término. Elaboración propia.

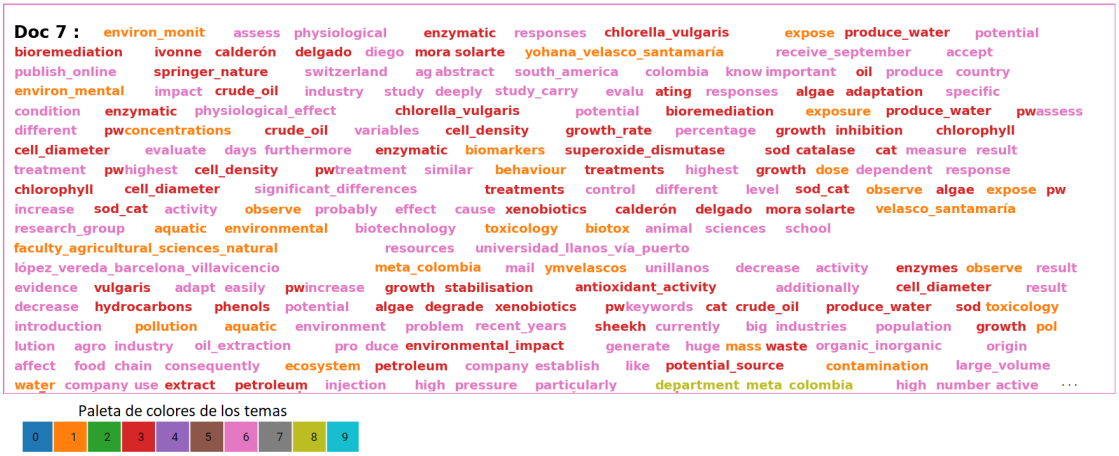


Figura 5-14.: Fragmento del documento 7 según con el tema 6 como dominante representado por el color rosado, así como el tema codificado por color de cada término. Elaboración propia.

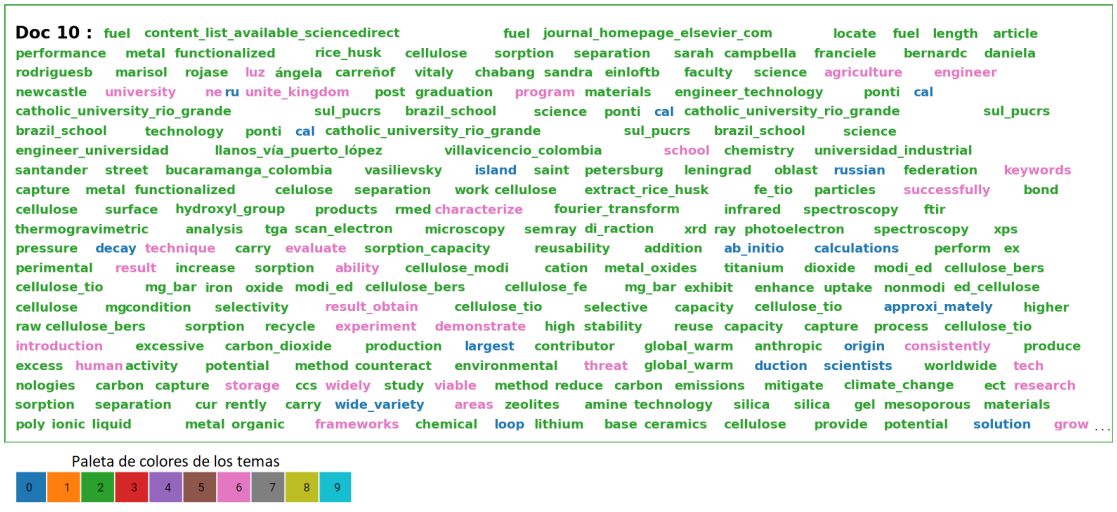


Figura 5-15.: Fragmento del documento 10 según con el tema 2 como dominante representado por el color verde, así como el tema codificado por color de cada término. Elaboración propia.

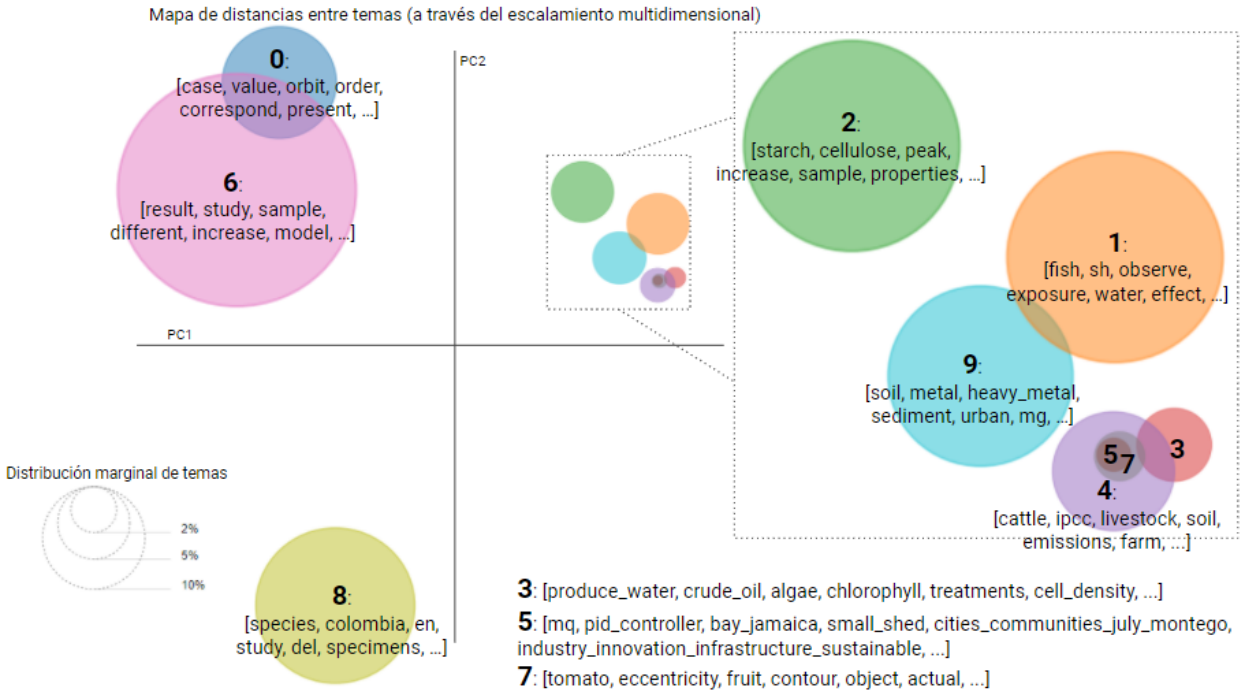


Figura 5-16.: Visualización de temas usando la representación 2D de los documentos con pyLDAvis. Elaboración propia.

dominantes para cada documento. Esta visualización se muestra en la figura **5-17**. Cada documento es expresado como una muestra acompañada del número de identificación del documento en el conjunto de datos cuya forma y color varía según su tema dominante. Los temas 3, 5 y 7 no se muestran debido a que ningún documento tiene asociado alguno de estos temas como dominante. Como se puede observar en la figura **5-17**, algunos documentos presentan una alta cercanía en la representación bidimensional usando el método t-SNE como se encierra en algunas regiones. En la figura **5-18** se amplían las tres regiones mencionadas anteriormente.

Las figuras **5-19**, **5-20** y **5-21** permiten detallar y comparar la razón de la similitud entre los documentos debido a la distribución de términos por tema en cada uno de los documentos, al usar la mayor probabilidad de los términos del documento a cada uno de los temas latentes. En la figura **5-19** y encerrados en la figura **5-17** con una circunferencia de color verde, mostrada igualmente en la subfigura **5-18**a, se comparan los documentos 121, 51 y 12, los cuales se acercaban más al tema 8 que al 6, no obstante, permaneciendo asignados al tema 6 por presentar mayor probabilidad. El título de estos documentos en conjunto con su probabilidad de pertenencia a dichos temas cercanos se muestran en la tabla **5-7**. En la figura **5-20**, los documentos 32 y 30 mostraron ser muy cercanos a los documentos del tema 2 aunque estaban asignados, por tener la probabilidad más alta, al tema 6, correspondientes a la circunferencia de color naranja mostrada en la subfigura **5-18**b. Adicionalmente, el título

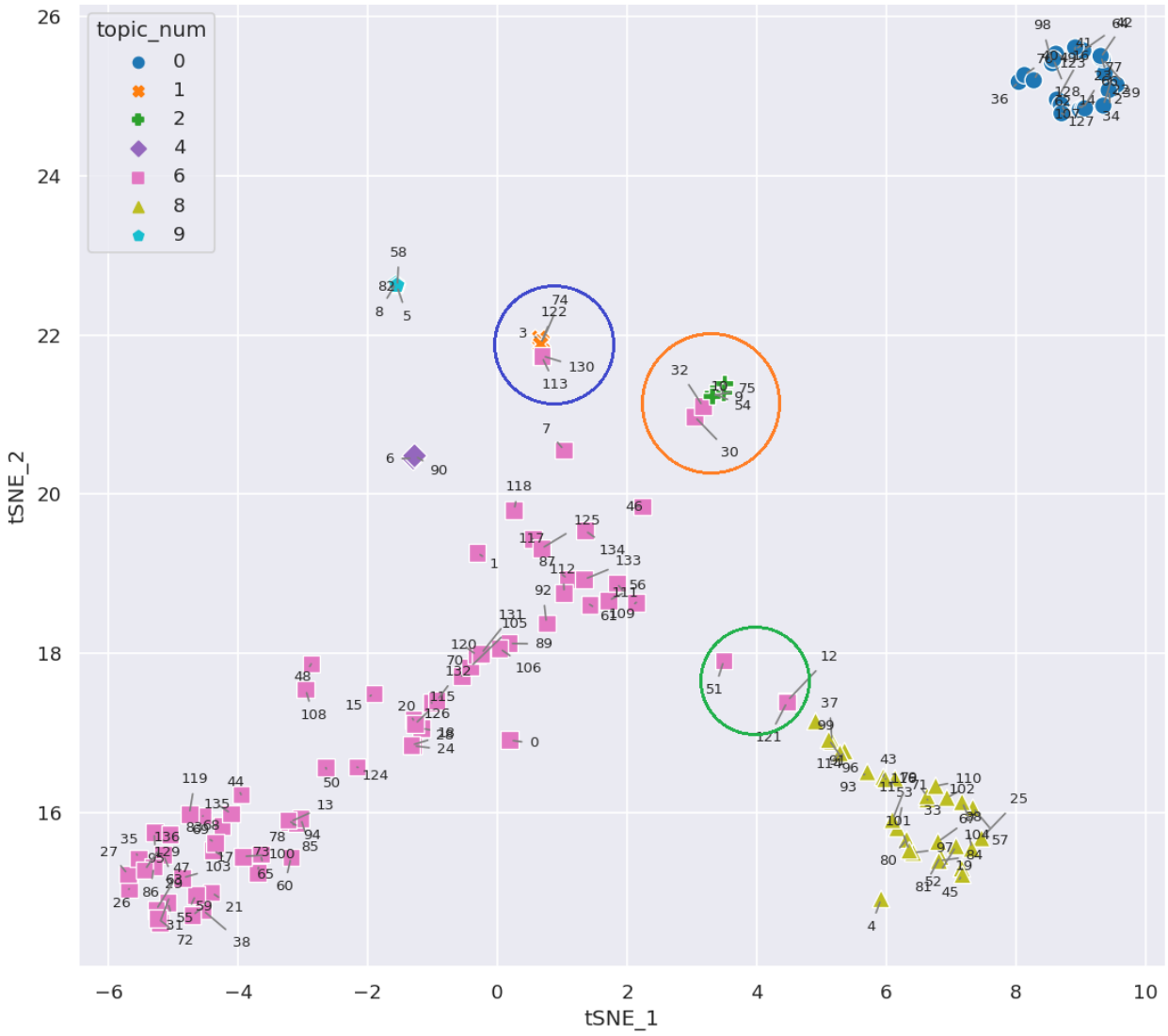
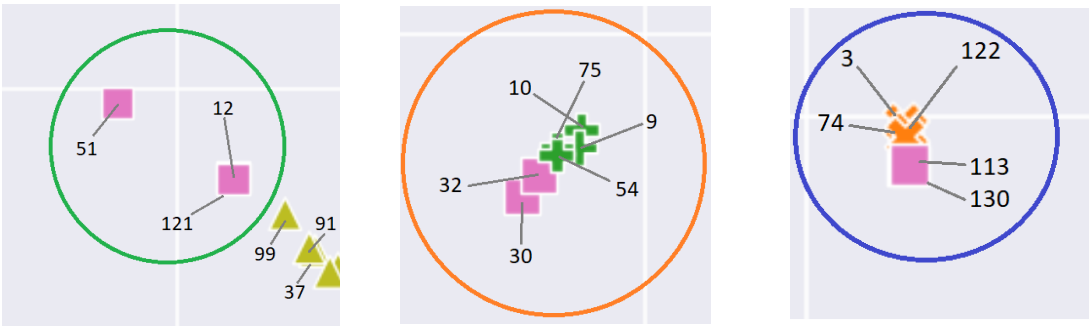


Figura 5-17.: Visualización en una representación bidimensional de los documentos con su identificador y tema dominante usando el método t-SNE. Elaboración propia.



(a) Docs. 12, 51 y 121 entre los temas 6 (rosado) y 8 (oliva).

(b) Docs. 32 y 30 entre los temas 6 (rosado) y 2 (verde).

(c) Docs. 113 y 130 entre los temas 6 (rosado) y 1 (naranja).

Figura 5-18.: Regiones con documentos cercanos o similares en la representación bidimensional usando el método t-SNE de acuerdo con la distribución de términos por tema dominante. Elaboración propia.

de estos documentos en conjunto con su probabilidad de pertenencia a dichos temas cercanos se muestran en la tabla **5-8**. Por último, en la figura **5-21**, se comparan los documentos 130 y 113 siendo igualmente muy cercanos al tema 1, a pesar de, tener la probabilidad más alta y por tanto, ser asignados al tema 6, estos documentos se encuentran encerrados por una circunferencia de color azul mostrada en la subfigura **5-18**c. Igualmente, el título de estos documentos en conjunto con su probabilidad de pertenencia a dichos temas cercanos se muestran en la tabla **5-9**.

Tabla 5-7.: Descripción documentos cercanos para los temas 6 y 8 en la representación bidimensional usando el método t-SNE de acuerdo con la distribución de términos por tema, según su pertenencia a los dos temas más probables.

ID Doc.	Título del documento	ID Tema dom. primario	Prob. Tema Dom.[1] primario	ID Tema dom. secundario	Prob. Tema Dom.[1] secundario
12	"*Health status of the elderly in life centers [Estado de saúde dos idosos dos centros de vida] [Estado de salud de los adultos mayores de los centros vida]*"	6	0.510	8	0.471
51	"*Origin and cross-century dynamics of an avian hybrid zone*"	6	0.589	8	0.367
121	"*Fish farming of native species in Colombia: Current situation and perspectives*"	6	0.486	8	0.453

[1] Probabilidad del tema en el documento

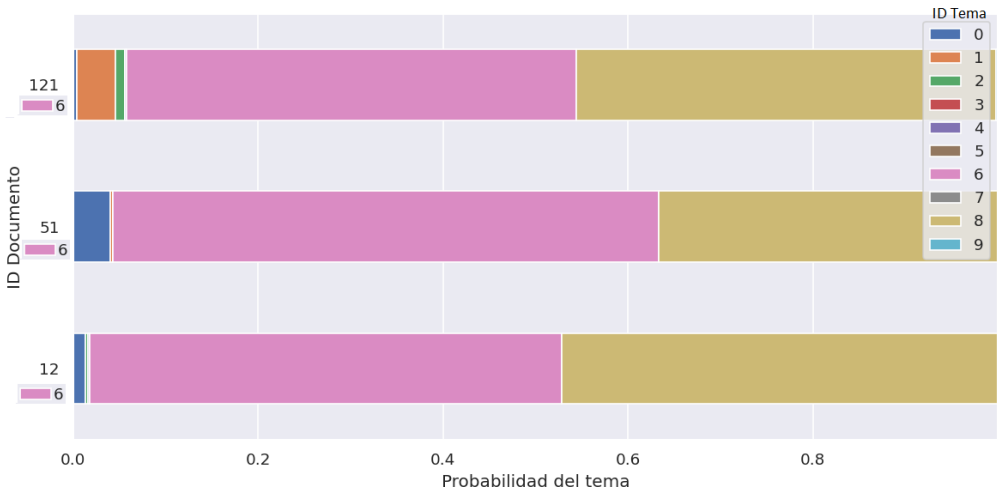


Figura 5-19.: Comparación entre documentos cercanos o similares para los temas 6 y 8 en la representación bidimensional usando el método t-SNE de acuerdo con la distribución de términos por tema, según su proporción de pertenencia a cada tema. Elaboración propia.

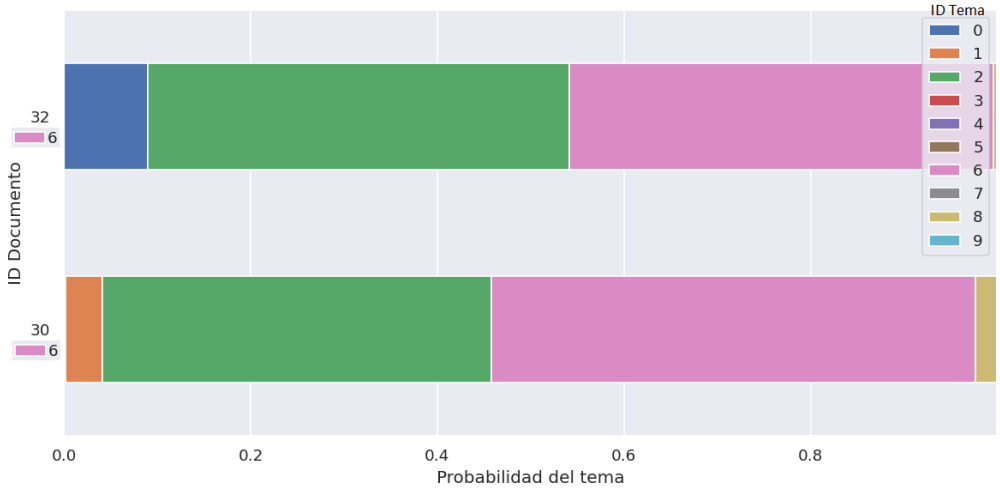


Figura 5-20.: Comparación entre documentos cercanos o similares para los temas 6 y 2 en la representación bidimensional usando el método t-SNE de acuerdo con la distribución de términos por tema, según su proporción de pertenencia a cada tema. Elaboración propia.

Tabla 5-8.: Descripción documentos cercanos para los temas 6 y 2 en la representación bidimensional usando el método t-SNE de acuerdo con la distribución de términos por tema, según su pertenencia a los dos temas más probables.

ID Doc.	Título del documento	ID Tema dom. primario	Prob. Tema Dom.[1] primario	ID Tema dom. secundario	Prob. Tema Dom.[1] secundario
30	"*Dynamics and use of nitrogen in biofloc technology - BFT*"	6	0.518	2	0.417
32	"*Effect of temperature and air equivalence ratio on energy potential of syngas produced from oil palm shells gasification*"	6	0.454	2	0.451

[1] Probabilidad del tema en el documento

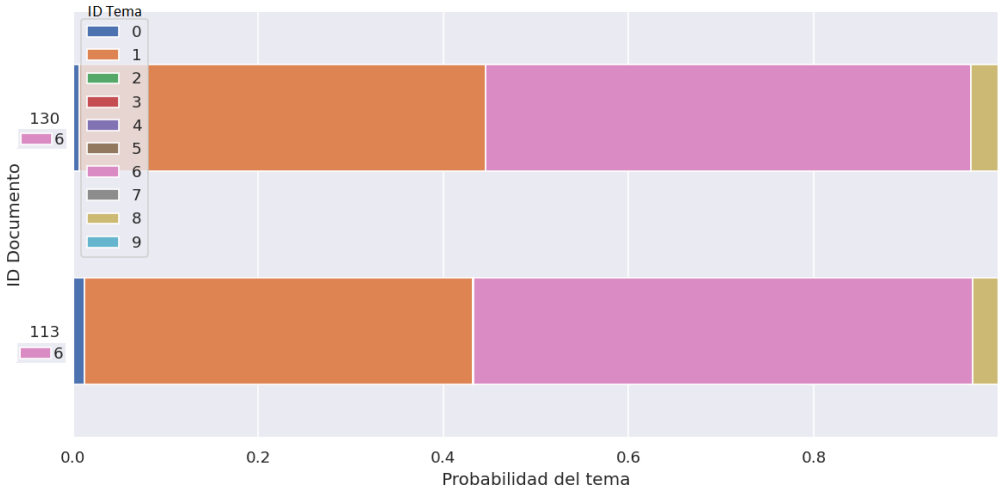


Figura 5-21.: Comparación entre documentos cercanos o similares para los temas 6 y 1 en la representación bidimensional usando el método t-SNE de acuerdo con la distribución de términos por tema, según su proporción de pertenencia a cada tema. Elaboración propia.

Tabla 5-9.: Descripción documentos cercanos para los temas 6 y 1 en la representación bidimensional usando el método t-SNE de acuerdo con la distribución de términos por tema, según su pertenencia a los dos temas más probables.

ID Doc.	Título del documento	ID Tema dom. primario	Prob. Tema Dom.[1] primario	ID Tema dom. secundario	Prob. Tema Dom.[1] secundario
113	"*Evidence of small modulation of ethinylestradiol induced effects by concurrent exposure to trenbolone in male eelpout Zoarces viviparus*"	6	0.539	1	0.419
130	"*Behavioural and gill histopathological effects of acute exposure to sodium chloride in moneda (Metynnis orinocensis)*"	6	0.523	1	0.439

[1] Probabilidad del tema en el documento

5.2. Evaluación cualitativa

5.2.1. Resultados

Los temas propuestos por los expertos que se obtuvieron de la valoración por inferencia temática (V1) son mostrados en la figura **5-22**, siendo agrupados por los temas latentes del modelo LDA presentados en el apartado "Temas latentes - LDA". Los temas propuestos son coloreados y ordenados según el nivel de experticia que cada experto consideró que tenía para inferir y proponer dichos temas, donde el nivel superior fue "Muy experto" con el color rojo, seguido de "Experto" con naranja, luego "Poco experto" con verde y por último, "Nada experto" con azul. En esta figura se observa que todos los temas latentes fueron evaluados por al menos un experto, siendo el tema latente 3 el que más temas propuestos tuvo y el tema latente 2 el que menos. Igualmente, todos los temas latentes tuvieron al menos una valoración con un nivel de experticia alto ("Experto" o "Muy experto"), a excepción del tema latente 2 con dos temas propuestos de nivel "Poco experto". En dicha figura, también se observa que en la mayoría de los casos existe una terminología técnica en los temas propuestos por parte de los expertos, esto ocurre principalmente en los niveles altos de experticia, como por ejemplo, en el tema 3 con el tema propuesto "Acuicultura" de un nivel "Muy experto" y con el tema propuesto "Tratamiento de agua" cuyo nivel de experticia es "Nada experto". Otro caso a destacar es en el tema latente 0 con los temas propuestos "Dinámica orbital" y "Sistemas de energía alternativos" siendo el primero más técnico y de nivel "Muy experto" y el segundo de un nivel "Nada experto". Finalmente, un caso particular se muestra en el tema latente 7 cuyos temas propuestos de nivel "Muy experto" son de mayor número de palabras que los niveles de experticia inferiores, siendo está tema latente el único que presentó este caso dado que en los otros temas latentes se tendía a disminuir la cantidad de palabras

usadas para construir el tema propuesto según aumentaba el nivel de experticia.

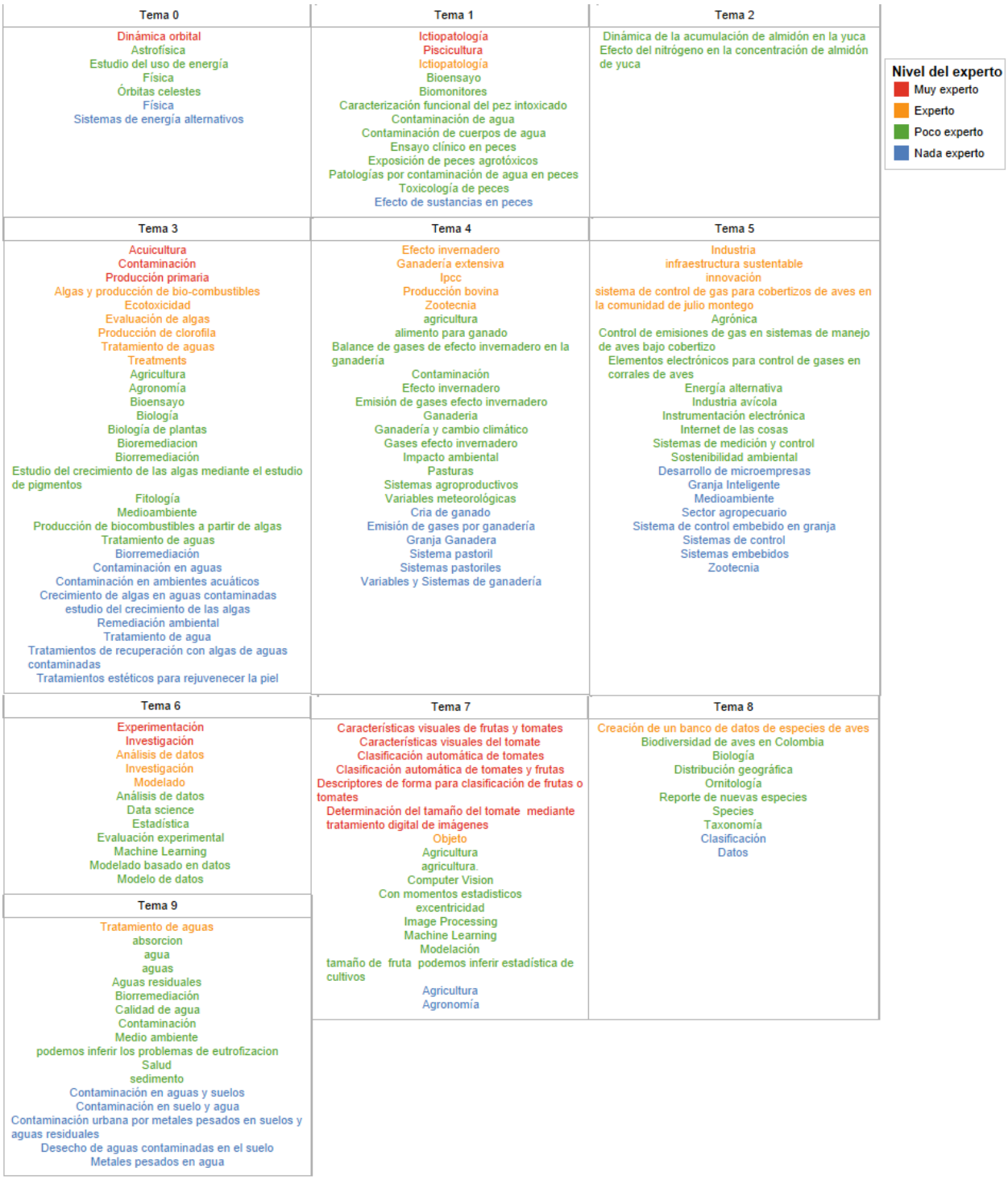


Figura 5-22.: Temas propuestos por los expertos según el tema y su nivel de experticia en inferencia temática (V1). Elaboración propia.

Posteriormente, para relacionar los términos de los temas propuestos por los expertos con otros términos en cada uno de los temas latentes se realizó un análisis de datos, en investigación cualitativa, aplicando el proceso de codificación de palabras. El proceso de codificación consistió en transformar cada palabra de forma independiente aplicando los siguientes pasos: *i)* transformación de plural a singular, *ii)* unificación de idioma (traducción de palabras al Español) y *iii)* eliminación manual de *stopwords* en Español. Con estas palabras codificadas se generaron nubes de palabras mostradas en la figura **5-23**. El tamaño de cada palabra es proporcional a la frecuencia de existencia de la misma. En dicha figura se observa como algunas palabras predominan notoriamente en comparación con otras del mismo tema en la mayoría de los casos, por ejemplo en los temas 0, 6 y 8 se destacan las palabras "física" y "energía", "dato" y "especie", respectivamente. Caso contrario, los temas 2 y 4 no muestran una diferencia destacable en los tamaños de las palabras.

Usando la misma representación de la figura **5-23** se clasificaron las palabras, obtenidas del proceso de codificación, en cuatro grupos cada uno correspondiente a un nivel de experticia, siendo "Muy experto", "Experto", "Poco experto" y "Nada experto" estas representaciones por grupos de experticia para cada tema latente se muestran en las figuras **5-24** y **5-25** abarcando desde el tema 0 hasta el 5 y desde el 6 hasta el 9, respectivamente. En dichas figuras se observa que varios niveles de experticia siguen presentando palabras de tamaños predominantes, casos como el tema 0 en el nivel "Nada experto" las palabras "física", "energía" y "sistema" tienen un mayor relevancia que "alternativo". Otro ejemplo, ocurre en el tema 5, donde las palabras "sistema" y "control" tienen un mayor tamaño tanto en el nivel "Poco experto" como en "Nada experto". Otro caso particular a destacar es en el tema 6, donde la palabra "dato" destaca notoriamente en los niveles "Experto" y "Poco experto". Por último, en el tema 9 la palabra "agua" hace presencia en los niveles "Experto", "Poco experto" y "Nada experto", destacando en los dos últimos. Por otra parte, se puede observar que los niveles bajos de experticia, como los son "Poco experto" y "Nada experto", de cada tema, es donde suelen haber mayor cantidad de palabras dada la terminología no tan técnica usada por los expertos requerida para construir los temas propuestos. No obstante, en el tema 7 ocurre todo lo contrario, siendo el nivel "Experto" el que mayor número de palabras tiene y "tomate" es la palabra predominante y por tanto, de mayor tamaño en dicho nivel.

Finalmente se estableció una comparación entre los temas latentes y algunas de las categorías establecidas por SCOPUS, siendo los campos de estas categorías: "Scopus Subarea" y "Scival Topic Prominence", respectivamente. Para ello se recopilaron las categorías de dichos campos para cada uno de lo documentos y se agruparon según el tema dominante en cada documento. En las figuras **5-26** y **5-27** se puede visualizar el contenido de las categorías más frecuentes por cada tema de acuerdo con los campos "Scopus Subarea" y "Scival Topic Prominence", respectivamente. Estas representaciones permitieron relacionar cada uno de los temas latentes con las categorías establecidas por SCOPUS más frecuentes.

Figura 5-23.: Nube de palabras obtenidas por el proceso de codificación de un análisis cualitativo a partir de los temas propuestos por expertos por cada tema latente. Elaboración propia.

Figura 5-24.: Nube de palabras obtenidas de un proceso de codificación para un análisis cualitativo de temas propuestos según nivel de experticia para cada tema latente. Parte 1 correspondiente a los 6 primeros temas del 0 al 5 de un total de 10 temas. Elaboración propia.

Figura 5-25.: Nube de palabras obtenidas de un proceso de codificación para un análisis cualitativo de temas propuestos según nivel de experticia para cada tema latente. Parte 2 correspondiente a los 4 últimos y restantes temas del 6 al 9 de un total de 10 temas. Elaboración propia.

Figura 5-26.: Relación de los temas latentes con las categorías asociadas de áreas por SCOPUS. Elaboración propia.

Respecto a las valoraciones de la asociatividad (V2) y la valoración por grado de representatividad (V3), se evaluaron un total 68 artículos científicos de 137 (aprox. 49.3 %) distribuidos como se muestra en la tabla **5-10**, la cual muestra la cantidad de documentos evaluados

Figura 5-27.: Relación de los temas latentes con las categorías asociadas de temas por Scival Topic Prominence. Elaboración propia.

por cada área del conocimiento frente a la cantidad existente de documentos asignados a dicha área que compone el conjunto de datos. En esta tabla se observa que el área con mayor cantidad de documentos evaluados en relación a la cantidad total de documentos es "Agricultural and Biological Sciences" en ambas valoraciones con 28 (49.1 %) y 27 (47.3 %) artículos científicos, respectivamente. Además, el área del conocimiento con mayor porcentaje de documentos en relación a los disponibles y asignados a la misma es "Environmental Science" con 11 documentos (64.7 %). Por otra parte, "Physics and Astronomy" y "Mathematics" son las áreas con menor cantidad de documentos evaluados en relación al total de documentos existentes con una cantidad de cuatro, no obstante, es "Physics and Astronomy" la que menor proporción de documentos evaluados tiene en relación a los asignados en dicha área con aproximadamente el 19.0 %.

Tabla 5-10.: Resumen cantidad de artículos científicos evaluados en Valoración de la asociatividad (V2) y Valoración por grado de representatividad (V3).

	Valoración		
Área	**Asociatividad (V2)**	**Representatividad (V3)**	**Documentos asignados**
"Agricultural and Biological Sciences"	28 (49.1 %)	27 (47.3 %)	57
"Engineering"	12 (50.0 %)	12 (50.0 %)	24
"Computer Science"	13 (59.1 %)	13 (59.1 %)	22
"Physics and Astronomy"	4 (19.0 %)	4 (19.0 %)	21
"Environmental Science"	11 (64.7 %)	11 (64.7 %)	17
"Biochemistry, Genetics and Molecular Biology"	7 (46.6 %)	7 (46.6 %)	15
"Mathematics"	4 (33.3 %)	4 (33.3 %)	12

En las tablas **5-11** y **5-12** se muestra en detalle la cantidad de evaluaciones por área del conocimiento según el grado de coherencia tanto en Valoración de la asociatividad (V2) y Valoración por grado de representatividad (V3). En estas tablas se observa que la mayor cantidad de evaluaciones realizadas fueron del área "Agricultural and Biological Sciences" en ambos casos con 32 evaluaciones. Siendo para el caso de V2 el nivel "Coherente" de esta misma área, el que obtuvo la mayor cantidad de evaluaciones con 18; mientras en V3 fue el nivel "Poco coherente" con 21 evaluaciones. Caso contrario, el área con menor cantidad fue "Physics and Astronomy" con 4 evaluaciones en cada valoración. En V2, esta área tuvo 2 evaluaciones en "Coherente" siendo el nivel con mayor cantidad de evaluaciones y en V3, las 4 evaluaciones existentes de esta área para esta valoración se concentraron en "Poco coherente". En términos generales, se evidencia que en la valoración V2 el nivel con mayor cantidad de evaluaciones fue "Coherente" con 40 (43.95 %) del total de evaluaciones realizadas en V2, mientras que en la valoración V3 fue "Poco coherente" con 50 (54.34 %) de las evaluaciones de V3. En total se obtuvieron 183 evaluaciones donde 91 correspondieron a V2 y 92 a V3. Clasificando los grados de coherencia en dos grupos "Altos" y "Bajos" se obtuvo un 53.84 % de evaluaciones altas en V2 y un 31.52 % en V3. La cantidad de evaluaciones por cada artículo científico para cada valoración, se puede observar en detalle en el anexo Anexo: Número de evaluaciones cualitativas por cada artículo científico (Anexo D) en las figuras **D-1** y **D-2**, respectivamente.

Tabla 5-11.: Detalle cantidad de evaluaciones por área del conocimiento según el grado de coherencia en Valoración de la asociatividad (V2).

	Grado de coherencia				
Área	**Muy Coherente**	**Coherente**	**Poco Coherente**	**Nada Coherente**	**TOTAL**
"Agricultural and Biological Sciences"	**3**	**18**	**7**	**4**	32
"Engineering"	**1**	**6**	**4**	**2**	13
"Computer Science"	**1**	**2**	**6**	**6**	15
"Physics and Astronomy"	**1**	**2**	**1**	0	4
"Environmental Science"	**3**	**7**	**3**	**1**	14
"Biochemistry, Genetics and Molecular Biology"	0	**5**	**2**	**1**	8
"Mathematics"	0	0	**4**	**1**	5
TOTAL	9	40	27	15	91

Tabla 5-12.: Detalle cantidad de evaluaciones por área del conocimiento según el grado de coherencia en Valoración por grado de representatividad (V3).

	Grado de coherencia				
Área	**Muy Coherente**	**Coherente**	**Poco Coherente**	**Nada Coherente**	**TOTAL**
"Agricultural and Biological Sciences"	**1**	**6**	**21**	**4**	32
"Engineering"	0	**6**	**4**	**3**	13
"Computer Science"	0	**1**	**12**	**2**	15
"Physics and Astronomy"	0	0	**4**	0	4
"Environmental Science"	**4**	**4**	**5**	**2**	15
"Biochemistry, Genetics and Molecular Biology"	0	**6**	**1**	**1**	8
"Mathematics"	0	**1**	**3**	**1**	5
TOTAL	**5**	**24**	**50**	**13**	92

Por último, en las tablas **5-13** y **5-14** se muestra en detalle la cantidad de evaluaciones por área del conocimiento según el tema dominante del documento tanto en Valoración de la asociatividad (V2) y Valoración por grado de representatividad (V3). En estas tabla se observa que el área "Agricultural and Biological Sciences" presenta evaluaciones a documentos asignados a diferentes temas dominantes siendo que la mayor cantidad de ellas se encuentran asociadas al tema 6 con 14 en ambas valoraciones. Por otra parte, el área "Physics and Astronomy" muestra una clara asignación de las 4 evaluaciones de documentos asociadas al tema 0, en ambas valoraciones.

Tabla 5-13.: Detalle cantidad de evaluaciones por área del conocimiento según el tema dominante del documento en Valoración de la asociatividad (V2).

	Tema dominante										
Área	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**TOTAL**
"Agricultural and Biological Sciences"	**1**	0	**1**	0	**2**	0	**14**	0	**12**	**2**	32
"Engineering"	**4**	0	0	0	0	0	**9**	0	0	0	13
"Computer Science"	**1**	0	0	0	0	0	**14**	0	0	0	15
"Physics and Astronomy"	**4**	0	0	0	0	0	0	0	0	0	4
"Environmental Science"	0	**2**	0	0	0	0	**6**	0	**4**	**2**	14
"Biochemistry, Genetics and Molecular Biology"	0	**1**	0	0	0	0	**3**	0	**4**	0	8
"Mathematics"	**4**	0	0	0	0	0	**1**	0	0	0	5
TOTAL	14	3	1	0	2	0	47	0	20	4	91

Tabla 5-14.: Detalle cantidad de evaluaciones por área del conocimiento según el tema dominante del documento en Valoración de la representatividad (V3).

	Tema dominante										
Área	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**TOTAL**
"Agricultural and Biological Sciences"	**1**	0	**1**	0	**1**	0	**14**	0	**13**	**2**	32
"Engineering"	**4**	0	0	0	0	0	**9**	0	0	0	13
"Computer Science"	**1**	0	0	0	0	0	**14**	0	0	0	15
"Physics and Astronomy"	**4**	0	0	0	0	0	0	0	0	0	4
"Environmental Science"	0	**2**	0	0	0	0	**6**	0	**5**	**2**	15
"Biochemistry, Genetics and Molecular Biology"	0	**1**	0	0	0	0	**3**	0	**4**	0	8
"Mathematics"	**4**	0	0	0	0	0	**1**	0	0	0	5
TOTAL	14	3	1	0	1	0	47	0	22	4	92

6. Discusión y análisis

La primera búsqueda sistemática de parámetros sirvió como base para interpretar el impacto de las diferentes configuraciones de las etapas de preprocesamiento, representaciones textuales y parametrización del modelo LDA. Esto permitió realizar una segunda búsqueda sistemática con menos parámetros descartando aquellos que desde la primera exploración sistemática mostraron no ser indispensables para el contexto del modelo y el conjunto de datos, permitiendo una exploración más detallada de los parámetros de mayor impacto para su análisis. De esta forma, se realizó la segunda búsqueda sistemática de parámetros, enfocada en comparar las etapas de "*stemming*", "creación de bigramas y trigramas" y "poda" en el preprocesamiento, la representación textual (i.e. BoW, TF-IDF, Binaria) y los parámetros del modelo LDA, definidos para la implementación en Gensim, como lo son "alpha", "eta" y número de temas. Los resultados de esta segunda búsqueda fueron resumidos brevemente en la tabla **5-3** para escoger la mejor configuración. Estos resultados, permitieron concluir que:

- Aplicar o no la etapa de "***Stemming***" parecía no tener mayor impacto. Sin embargo, se observó que en los modelos con pocos temas ($K = 10$) se desempeñaban un poco mejor cuando no se aplicaba. Considerando facilitar la interpretabilidad, se determinó fijar un número bajo para el número de temas. Por tal motivo, se optó por omitir esta etapa en el preprocesamiento.

- En la etapa "Creación de Bigramas y Trigramas" se observó que el impacto de ésta no era considerable, principalmente en los modelos con pocos temas. Adicionalmente, el propósito de esta etapa era que el diccionario incluyera términos de palabras compuestas, comunes en artículos científicos, como por ejemplo "statistical_moments" y "academy_sciences". Por lo cual, se optó por incluir esta etapa.

- En la etapa "Poda", los desempeños tuvieron valores similares, destacando que la mayor dispersión se presentó en modelos con pocos temas. No obstante, los desempeños de aquellos modelos sin implementar esta etapa resultaron ser levemente mejor, en términos generales. Por tal razón, se optó por omitir esta etapa.

- La representación textual con mejor de desempeño fue TF-IDF, no obstante, también presentó una mayor dispersión. Por otra parte, BoW resultó más estable en comparación con TF-IDF con una menor dispersión aunque sin alcanzar valores comparables en promedio por muy poco. Finalmente, la representación binaria, mostró resultados

muy inferiores a sus contrapartes, destacando únicamente la baja dispersión. Por tal motivo, se optó por implementar BoW como representación textual.

- En el parámetro "**alpha**", se observó que, en la mayoría de los casos, se desempeñó mejor usar el valor de "auto" como parámetro en modelos con pocos temas latentes. Similar ocurrió para el parámetro "**eta**" con desempeños un poco mejores con la configuración "symmetric".

- En el número de temas (**num_topics**), se pretendía desde un inicio establecer un número bajo de temas debido a la poca cantidad de artículos científicos procesados (137) en el conjunto de datos obtenido, buscando así poder generalizar y agrupar los documentos en pocos temas. Sin embargo, era necesario corroborar su impacto en el modelo para el conjunto de datos. Fue evidente la tendencia a disminuir el desempeño con el aumento del número de temas. Por tal motivo, se confirmó la idea inicial de establecer el número de temas en 10.

Aplicando la mejor configuración para el entrenamiento del modelo LDA, se obtuvieron 10 temas latentes mostrados en la tabla **5-4** y la figura **5-11**, con sus 10 términos más probables para cada tema latente. El conjunto de datos que se implementó está conformado por 137 artículos científicos con un tamaños que oscilan entre los 100 y 4,700 términos aproximadamente.

Se interpretaron de manera conjunta los resultados cuantitativos y cualitativos para inferir y etiquetar un tema a cada uno de los 10 temas latentes obtenidos por el modelo LDA. Para ello se relacionaron los temas propuestos por los expertos para cada tema latente ponderando el nivel de experticia señalado por los mismos, además de considerar el grado de generalización de los temas propuestos. Es decir, se dio más relevancia a los temas propuestos de expertos que se consideraron de mayor experticia cuando hubiera diversidad o contraste en dichos temas como por ejemplo en el tema 6 con los temas propuestos "Investigación" y "Experimentación" de un nivel "Muy experto" frente a los términos de temas propuestos "Modelado" y "Dato" de nivel "Poco experto", siendo incluso "Dato" el término más usado en los temas propuestos para este tema latente. Otro ejemplo respecto a la generalización y uso de tecnicismos, se observó en el tema 1, donde los expertos propusieron el tema "Ictiopatología"[1] en los niveles "Muy experto" y "Experto" de experticia frente a los términos usados en el nivel "Poco experto" cuya terminología está igualmente relacionada pero no es tan técnica, como por ejemplo "Pez", "Contaminación" y "Agua".

Los identificadores de temas y sus descripciones de los temas latentes correspondientes son mostrados en la tabla **6-1**. En algunos casos, se consideró que los temas latentes podrían abarcar más de un tema relacionado, por tal motivo, se adicionaron las descripciones de

[1] La Ictiopatología es el estudio de las enfermedades de los peces.

temas que mejor explicaban el contenido de los temas latentes de acuerdo con sus términos más probables, cada uno separados con el caracter "|".

Tabla 6-1.: Inferencia y descripción de temas latentes, obtenidos del modelo LDA, por parte de los autores a partir de temas propuestos por expertos.

Id. Tema	Descripción del tema latente	Descripción en Inglés
0	Física | Órbitas celestes | Energía	Physics | Celestial orbits | Energy
1	Ictiopatología	Ichthyopathology
2	Estudio | Almidón de yuca	Study | Cassava starch
3	Acuicultura | Contaminación del agua	Aquaculture | Water contamination
4	Efecto Invernadero | Ganadería	Greenhouse effect | Livestock
5	Industria | Avicultura | Innovación tecnológica	Industry | Poultry | Technological innovation
6	Experimentación | Investigación	Experimentation | Research
7	Tratamiento digital de imágenes | Agricultura	Digital image processing | Agriculture
8	Ornitología	Ornithology
9	Contaminación en aguas y suelos	Water and soil contamination

A pesar tener los valores de coherencia más altos los temas 3, 5 y 7, no fueron considerados por ningún artículo científico como su tema dominante, sino temas asociados a otros temas dominantes entre los artículos del conjunto de datos. Al contrario que los temas con valor de coherencia bajos como el 6 y el 8, siendo los temas dominantes de una mayor cantidad de documentos. Esto puede ser consecuencia de los fenómenos de sobreajuste (*overfitting)* y subajuste (*underfitting*). Se considera que es sobreajuste en los temas 3, 5 y 7 dada la terminología mayormente técnica o puntual que conforman dichos temas, como por ejemplo, "produce_water" y "crude_oil", "pid_controller" y "arduino" y "tomato" y "pixels", respectivamente. Dicha terminología restringe la generalización o abstracción objetivo en el modelado de temas debido a la poca cantidad de documentos y la variabilidad temática que la conforman. Por otra parte, se considera que los temas 6 y 8 pueden presentar subajuste debido a una terminología más general y prácticamente transversal en el contexto académico y científico de los documentos. Términos como "result" y "study", se consideran transversales en la investigación que es la razón de escritura de artículos científicos, y "species" y "colombia" siendo términos genéricos al contexto de los documentos debido al foco principal de investigación (sector agropecuario) y la región objetivo (región Orinoquía, Colombia) de la Universidad de los Llanos.

Partiendo del hecho que un artículo científico está conformado por términos que, si se analizan de manera conjunta, algunos de ellos resultan ser más representativos para otros temas que para el tema dominante del documento, como se muestra en la figura **5-12** y en los fragmentos de los documentos de las figuras **5-13** al **5-15**. Factores como el tecnicismo, la

sinonimia, condición de sinónimo que según la RAE[2]: "*Dicho de una palabra o de una expresión: Que, respecto de otra, tiene el mismo significado o muy parecido, como 'empezar' y 'comenzar'*"[3], y la polisemia, según la RAE: "*Pluralidad de significados de una expresión lingüística*"[4] como 'banco' y 'lengua', pueden influir en que un documento tenga cierta proporción de temas en su contenido.

Para visualizar la distribución de los artículos científicos por temas dominantes y su relación de similitud entre ellos, se generó una visualización bidimensional implementando el método t-SNE (subsección 4.3.2). Esta visualización de la figura **5-17**, permitió comprobar que la asignación de temas dominantes era apropiada, al considerar que los grupos estaban bien definidos y no se encontraban documentos atípicos. Un fenómeno que se observó fue que ciertos documentos parecían ser parte más a un tema que al que estaban asignados. Por tal motivo, se comprobó que dichos documentos, presentaban ese comportamiento por que tenían probabilidades similares entre los temas en cuestión. Por ejemplo, el documento 12 posee probabilidades similares para los temas 6 y 8. Tomando en cuenta el nombre del documento, a pesar que esto no determine explícitamente el contenido del mismo, sirve como punto de referencia para comparar y analizar dicha relación, el cual es "*Health status of the elderly in life centers [Estado de saúde dos idosos dos centros de vida] [Estado de salud de los adultos mayores de los centros vida]*". Se puede observar que no hay una estrecha relación entre el documento y los temas 6 "Experimentación | Investigación" y 8 "Ornitología", de hecho se asume que con el tema 8 no tiene ninguna relación. Otro caso fue el artículo 32 que lleva el nombre "*Effect of temperature and air equivalence ratio on energy potential of syngas produced from oil palm shells gasification*" y presentó probabilidades entre los temas 6, 2 "Estudio | Almidón de yuca" y 0 "Física | Órbitas celestes | Energía", en este caso y según el título, se observa que el documento presenta un estudio o investigación, relacionado al tema 6, de elementos energéticos, presente en el tema 0, en elementos del sector agrícola, asociado al tema 2. Esto permite deducir que este artículo científico tiene relación con dichos temas. Adicionalmente y de igual forma, se analizó cada documento que presentó este comportamiento permitiendo interpretar que su contenido temático es correcto o que no fue posible relacionarlo estrechamente con algún tema, esto último puede ser posible por el alto grado de especificidad del tema del documento, presentándose esto en pocos casos, como el ejemplo mostrado anteriormente del documento 32.

Es evidente la predominancia en la mayoría de documentos hacia el tema 6 "Experimentación | Investigación", a pesar de tener un valor bajo de la medida de coherencia. Esto es importante y destacable por el significado del tema en sí y del contexto del conjunto de datos. El conjunto de datos está conformado por artículos científicos generadores de nuevo

[2]Real Academia Española
[3]Sinonimia: `https://dle.rae.es/sinonimia`
[4]Polisemia: `https://dle.rae.es/polisemia`

conocimiento cuyo contenido se basa en estudios, experimentación e investigación de diferentes áreas de conocimiento, es decir el tema 6.

Gracias a las evaluaciones realizadas por los expertos fue posible etiquetar y comparar cada uno de los temas. Basado en los temas propuestos y ponderando cualitativamente el nivel de experticia se construyeron las etiquetas que permitieron representar de mejor manera la interpretación semántica de los temas obtenidos por el modelo LDA implementado. Las evaluaciones permitieron obtener temas propuestos de nivel alto en todos los temas, a excepción del tema 2 "Estudio | Almidón de yuca", conformado por términos como "starch" ("almidón"), "cellulose" ("celulosa") y "peak" ("pico"), siendo los tres términos más probables para este tema, y cuyas palabras dominantes según las interpretaciones de los expertos fueron "almidón", "concentración" y "yuca", pertenecientes al nivel de experticia "Poco experto". Esto puede deberse al criterio de selección del tema implementado en el instrumento "whatTopic" (apartado Valoración por inferencia temática (V1) subsección 4.4.2) y a la heterogeneidad en las áreas de los expertos quienes evaluaron éste tema. Esto también se ve reflejado en el alcance del análisis cualitativo de los resultados debido a que la naturaleza del trabajo y la cantidad de investigadores expertos en relación a los artículos presentados aleatoriamente requiere una mayor cantidad de evaluaciones por documento o tema para aplicar métodos cualitativos como la evaluación interjueces. Por lo cual, como trabajo futuro se requiere incorporar investigadores evaluadores adicionales tanto internos como externos a la institución y hacer el análisis sobre una muestra aleatoria, o por conveniencia más pequeña, al conjunto de datos total que sea representativa para un análisis más detallado y profundo.

Por tal motivo, se aplicó un análisis cualitativo por medio de un proceso de codificación que consistió en transformar los términos usados por los expertos en los temas inferidos en la evaluación Valoración por inferencia temática (V1) para comparar y encontrar relaciones de conceptos comunes o ideas. En otras palabras, el proceso de codificación consistió en transformar cada término a una forma generalizada de su concepto que permitiera agrupar todos los demás términos cuya transformación fuera igual. En la figura **5-23** se mostró un conjunto de nubes de palabras referentes a los términos transformados según cada tema. Allí se observó que algunas palabras predominaron notoriamente en comparación con otras del mismo tema en la mayoría de los casos, por ejemplo en los temas 0, 6 y 8 se destacan las palabras "física" y "energía", "dato" y "especie", respectivamente. Esto permite interpretar que dichas palabras fueron claves y claramente importantes para los expertos en el proceso de inferencia de los temas propuestos. Caso contrario, los temas 2 y 4 no muestran una diferencia destacable en los tamaños de las palabras. Esto puede significar que *i)* no hubo términos destacables en dichos temas latentes para los expertos que les permitieran inferir las palabras relacionadas, o *ii)* el conocimiento técnico que los expertos no era suficiente para generalizar el tema latente y usar palabras que mejor lo conceptualizaran.

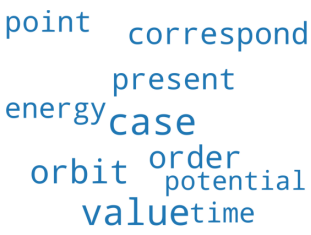
Posteriormente, se elaboró una visualización de palabras transformadas según su tema y nivel de experticia. Dicha visualización es mostrada en las figuras **5-24** (temas del 0 al 5) y **5-25** (temas del 6 al 9). Esta visualización permitió observar que no había una tendencia general en el uso de las palabras en todos los niveles de experticia, en su lugar, hubieron casos particulares como los siguientes: *i)* Ausencia de palabras, debido a que no se obtuvo una evaluación con dicho nivel de experticia, por ejemplo en el tema 2 en los niveles "Muy experto", "Experto" y "Nada experto". *ii)* Habían pocas palabras o una sola en el mismo nivel de experticia que incapacitaba un proceso de comparación, por ejemplo en el tema 0 en el nivel "Muy experto" con únicamente dos palabras ("dinámica" y "orbital") y en el tema 1 en el nivel "Experto" con sólo la palabra "Ictiopatología". *iii)* Habían varias palabras en un mismo nivel de experticia pero sin destacar ninguna entre las demás, por ejemplo en el tema 0 en el nivel "Poco experto" y en el tema 3 en "Muy experto". Por último, *iv)* Habían varias palabras en un mismo nivel de experticia con algunas de ellas claramente destacables, por ejemplo en el tema 1 en el nivel "Poco experto" con las palabras "pez", contaminación" y "agua"; y en el tema 6 en el nivel "Poco experto" destacando las palabras "dato", "modelado" y "análisis". Adicionalmente se observó que, de manera general, los niveles bajos de experticia de cada tema, como los son "Poco experto" y "Nada experto", es donde existían una mayor cantidad de palabras, se asume que es debido a la terminología no tan técnica usada por los expertos requerida para construir los temas propuestos.

Los artículos científicos que fueron parte de la evaluación realizada por las valoraciones de asociatividad (V2) y por grado de representatividad (V3) fueron detallados en las tablas **5-10**-**5-14**. Tomando en cuenta que, la valoración de asociatividad (V2) consistió en identificar un conjunto de términos asociados a un artículo científico mostrado y su nivel de coherencia y que, la valoración por grado de representatividad consistió en relacionar un conjunto de términos y su grado de representatividad ordinal a un artículo mostrado al igual que su nivel de coherencia. En total se evaluaron 68 documentos de 137 (aprox. 49.3 %) en cada una de las valoraciones V2 y V3. Respecto a la valoración por asociatividad (V2), el área del conocimiento con mayor cantidad de evaluaciones fue "Agricultural and Biological Sciences" con 32 (aprox. 35.1 %) de los cuales 18 tuvieron un nivel "Coherente" siendo el grado de coherencia con mayor número de evaluaciones. El área "Physics and Astronomy" con cuatro evaluaciones (aprox. 4.3 %) fue el área con menos documentos evaluados, de los cuales el nivel "Coherente" fue el nivel con más evaluaciones siendo 2. En cuanto a la valoración por grado de representatividad (V3), el área "Agricultural and Biological Sciences" fue la que mayor evaluaciones tuvo para un total de 32 (aprox. 34.7 %), de los cuales, 21 evaluaciones pertenecen al nivel "Poco coherente" siendo éste el nivel con mayor número de evaluaciones. El área con menor cantidad de evaluaciones fue "Physics and Astronomy" con cuatro (aprox. 4.3 %), lo cuales corresponden al nivel "Poco coherente".

En términos generales, el rango del porcentaje de los documentos evaluados por área fue del 19.0 % ("Physics and Astronomy") al 64.7 % ("Environmental Science") con una media aproximada de 46 % de evaluación a los documentos disponibles por área. Tomando en cuenta la cantidad de documentos del conjunto de datos (137) y la cantidad de expertos (31) que realizaron el proceso de evaluación, se considera un balance aceptable que permite interpretar el contenido temático e investigativo de los artículos científicos y relacionarlo con las evaluaciones hechas por los expertos. Adicionalmente, el 53.8 % y 31.5 % de las evaluaciones fueron de nivel alto en valoración por asociatividad (V2) y valoración por grado de representatividad (V3), respectivamente. Por tanto, se concluye que la metodología implementada para la construcción de la valoración por asociatividad (V2) tuvo un desempeño aceptable tomando como medida de desempeño la cantidad de evaluaciones que tuvieron un nivel alto. En otras palabras, los términos mostrados a los expertos en el instrumento "whatTopic" permitieron inferir, en su mayoría, un tema por parte de los expertos con un nivel alto de coherencia. Por otra parte, la valoración por grado de representatividad (V3), se considera no tuvo un desempeño aceptable por lo que los términos que se mostraron en el instrumento no permitieron representar de forma ordinal los documentos mostrados en esta valoración. Por tal motivo, se considera para trabajo futuro mejorar el criterio de selección y consolidación del conjunto de términos mostrados en el instrumento "whatTopic" en ambas valoraciones, principalmente, en la valoración por grado de representatividad (V3).

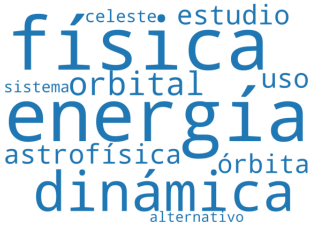
Finalmente, en las figuras **6-1**, **6-2** y **6-3** se observa un conjunto de nubes de términos referentes a temas de ejemplo como lo son 0, 3 y 6, respectivamente, que permitieron comparar y analizar las relaciones existentes entre los términos resultantes del modelo LDA, las descripciones propuestas por expertos tomando en cuenta su nivel de experticia, la relevancia de los términos resultantes al proceso de codificación y la relevancia entre las descripciones asociadas tanto en "Scopus Subarea" como "Scival Topic Prominence" de SCOPUS. Estos temas fueron seleccionados como ejemplos por que su comportamiento permitió resaltar características como: un balance entre el valor de coherencia, temas propuestos y dominancia de temas en el conjunto de datos (tema 0), un valor de coherencia alto pero ausencia total de dominancia de temas y por ende incapacidad de relacionar ciertas características como las descripciones de "Scopus Subarea" y "Scival Topic Prominence" (tema 3) y un valor de coherencia bajo pero con alta dominancia y gran cantidad de descripciones que facilitaron su análisis (tema 6).

A partir de las relaciones observadas se propuso asociar las descripciones propuestas con las categorías de los campos "Scopus Subarea" y "Scival Topic Prominence" predominantes. Estas relaciones son mostradas en la tabla **6-2**. Tomando en cuenta que las descripciones en "Scopus Subarea" y "Scival Topic Prominence" originalmente están en Inglés, se procedió a relacionar las descripciones propuestas en dicho idioma y así unificar el idioma para un análisis con menos modificaciones posibles. La asignación y comparación de estos campos está

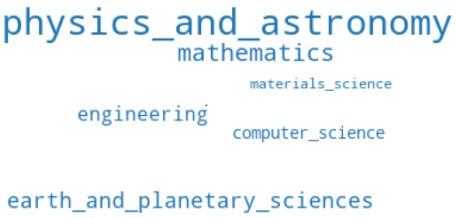


(a) Resultantes del modelo LDA.

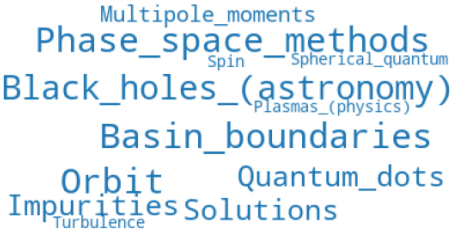
(b) Descripciones por expertos según nivel de experticia.



(c) Resultantes del proceso de codificación.

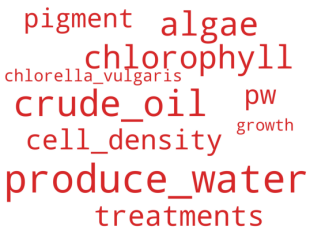


(d) Descripciones “Scopus Subarea”.



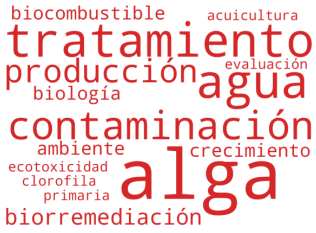
(e) Descripciones “Scival Topic Prominence”.

Figura 6-1.: Comparativo entre nubes de términos del tema 0 “Física | Órbitas celestes | Energía” ($CM_0 = 0.601$). Elaboración propia.



(a) Resultantes del modelo LDA.

(b) Descripciones por expertos según nivel de experticia.

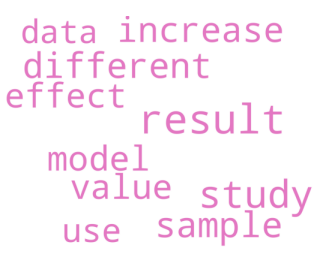


(c) Resultantes del proceso de codificación.

(d) Descripciones “Scopus Subarea”.

(e) Descripciones “Scival Topic Prominence”.

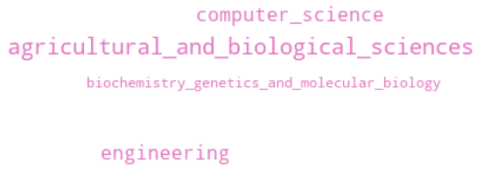
Figura 6-2.: Comparativo entre nubes de términos del tema 3 “Acuicultura | Contaminación del agua” ($CM_3 = 0.895$). Elaboración propia.



(a) Resultantes del modelo LDA.

(b) Descripciones por expertos según nivel de experticia.

(c) Resultantes del proceso de codificación.



(d) Descripciones “Scopus Subarea”.

(e) Descripciones “Scival Topic Prominence”.

Figura 6-3.: Comparativo entre nubes de términos del tema 6 “Experimentación | Investigación” ($CM_6 = 0.380$). Elaboración propia.

basada en los documentos y sus asociaciones preestablecidas por SCOPUS en el conjunto de datos, asignando así dichos campos un determinado tema dominante. Existen los temas 3, 5 y 7 que no tienen documento alguno, como bien se había mostrado con anterioridad. Por tal motivo, no fue posible asociar algún campo y en estos casos se agrega el término “Ninguno” para indicarlo. En algunos casos no hay etiquetas destacables en comparación con el resto por lo que se agregaron más de una. Esto se asume, es por la variabilidad temática y el margen de error del modelo entrenado LDA.

Si bien “Scopus subarea” y “Scival Topic Prominence” son descripciones que permiten asociar el tema o idea contenido en los documentos alojados en SCOPUS, se observó que “Scopus subarea” implementa una terminología general, incluso, a un nivel de áreas del conocimiento más no permite asociar un tema de manera específico. Sin embargo, “Scival Topic Prominence” propone temas más puntuales y asociados al contenido de los documentos usando una terminología más técnica, no obstante, en algunos casos el nivel de dichos temas se consideró aún superficial. En términos generales, se observó una gran relación entre las tres descripciones, estando más estrechamente relacionadas las descripciones propuestas y las definidas por “Scival Topic Prominence”. Esto puede deberse a la terminología técnica usada por los expertos para inferir los temas propuestos basados en el amplio conocimiento y su nivel de experticia en dichas áreas.

Tabla 6-2.: Comparación de descripciones de temas identificadas en la evaluación cualitativa y las categorías de temas: Propias (en Inglés), "Scopus Subarea" y "Scival Topic Prominence".

Tema		Descripciones más relacionadas
0	Descripción propuesta	"Physics | Celestial orbits | Energy"
	Scopus Subarea	"Physics and astronomy"
	Scival T.P.	"Black holes" y "Orbit"
1	Descripción propuesta	"Ichthyopathology"
	Scopus Subarea	"Environmental science"
	Scival T.P.	"Endolsufan sulfate", "Genotoxicity" y "Pesticides"
2	Descripción propuesta	"Study | Cassava starch"
	Scopus Subarea	"Chemistry"
	Scival T.P.	"Cycloaddition", "Adsorption", "Cassava" y "Starch"
3	Descripción propuesta	"Aquaculture | Water contamination"
	Scopus Subarea	(ninguno)
	Scival T.P.	(ninguno)
4	Descripción propuesta	"Greenhouse effect | Livestock"
	Scopus Subarea	"Agricultural and biological sciences"
	Scival T.P.	"Crude protein", "Tithonia diversifolia" y "Silvopastoral systems"
5	Descripción propuesta	"Industry | Poultry | Technological innovation"
	Scopus Subarea	(ninguno)
	Scival T.P.	(ninguno)
6	Descripción propuesta	"Experimentation | Research"
	Scopus Subarea	"Agricultural and biological sciences", "Engineering" y "Computer Science"
	Scival T.P.	"Medical imaging", "Cognitive radio", "Pathology" y "Histopathological images"
7	Descripción propuesta	"Digital image processing | Agriculture"
	Scopus Subarea	(ninguno)
	Scival T.P.	(ninguno)
8	Descripción propuesta	"Ornithology"
	Scopus Subarea	"Agricultural and biological sciences"
	Scival T.P.	"Colombia", "Bird species" y "Nests"
9	Descripción propuesta	"Water and soil contamination"
	Scopus Subarea	"Environmental science"
	Scival T.P.	"Soil", "Soil maps", "Road runoff" y "Suspended solids"

7. Conclusiones y trabajo futuro

7.1. Conclusiones

Se implementaron satisfactoriamente procesos metodológicos y evaluativos para la aplicación, ejecución y análisis de un método de aprendizaje computacional del estado del arte para el análisis automático de contenidos textuales enfocado en la tarea de modelado de temas, usando el método LDA, cuyo resultado permite ser de apoyo en el análisis documental y la gestión de conocimiento en instituciones de educación superior como la Universidad de los Llanos.

Se recopiló y consolidó un conjunto de datos de artículos científicos con el potencial de crecimiento no sólo para una institución sino varias de la región de la orinoquía colombiana, incluyendo diferentes versiones según etapas de preprocesamiento y representaciones textuales. Consecuentemente, se implementó el método computacional de LDA para el procesamiento y análisis de artículos científicos para la detección automática de temas latentes permitiendo así promover y desarrollar la adopción tecnológica de la región.

Se aplicaron dos metodologías para evaluar el desempeño del método LDA, tanto cuantitativa como cualitativa, obteniendo resultados que permitieron asociar los 10 temas latentes obtenidos por el modelo LDA con un valor de coherencia global de 0.639 con temas y descripciones propuestos por expertos en la evaluación cualitativa, como por ejemplo el tema 6 cuyos términos top-10 obtenidos por el modelo LDA son “result” (“resultado”), “study” (“estudio/estudiar”), “sample” (“muestra”), “different” (“diferente”), “increase” (“aumento/aumentar”), “model” (“modelo/modelar”), “value” (“valor/valorar”), “effect” (“efecto”), “use” (“uso/usar”) y “data” (“dato”) se asoció con la descripción “Experimentación | Investigación”. Llama la atención la ocurrencia, relevancia y transversalidad del tema 6 “Experimentación | Investigación” sobre el conjunto de datos, dada la naturaleza de los documentos siendo estos artículos científicos cuyo contenido se centra en investigación, experimentación y estudios científicos de diferentes áreas del conocimiento.

La evaluación cualitativa permitió obtener una interpretación con alto nivel de experticia, en la mayoría de temas, por parte de los expertos. Algunos casos, como los temas 4, 5, 8 y 9, presentaron ausencia de evaluaciones en al menos uno de los niveles altos de experticia y llegando, incluso, a que en el tema 2 no hubiera evaluaciones de alto nivel de experticia.

Se asume que se debe a la heterogeneidad en las áreas y niveles del conocimiento de los expertos. Igualmente, se ve reflejado en el alcance del análisis cualitativo de los resultados debido a que la naturaleza del trabajo y la cantidad de investigadores expertos, en relación a los artículos presentados aleatoriamente, requieren una mayor cantidad de evaluaciones por documento o tema para aplicar métodos cualitativos como la evaluación interjueces. De igual forma, según como se planteó la metodología de evaluación cualitativa el alcance no permitía tener una mayor cantidad de evaluaciones por documento o tema. Sin embargo, permitió un análisis primario del estado académico y científico del conjunto de datos procesado que sirve como base con potencial crecimiento y trabajos futuros.

Por tal motivo, se aplicó un análisis cualitativo por proceso de codificación que permitió comparar y encontrar relaciones entre los temas latentes resultantes del modelo LDA implementado y las categorías de campos preestablecidas por SCOPUS, como lo son "Scopus subarea" y "Scival Topic Prominence". Esto mostró que los temas latentes y las descripciones de los expertos estaban más relacionados con "Scival Topic Prominence" por presentar un mayor grado de especificidad técnica de los temas y la terminología usada para describirlos que "Scopus Subarea". Esto permitió conocer y comparar en un nivel más detallado los temas técnicos asociados a cada documento. Para algunos artículos científicos se encontraron temas que no existían en las categorías de Scopus ("Scopus Subarea" y "Scival Topic Prominence") o al menos, a un nivel detalle técnico que permitiera conocer a priori el tema del contenido del documento. Por ejemplo, el documento con ID 32 cuyo título es "*Effect of temperature and air equivalence ratio on energy potential of syngas produced from oil palm shells gasification*" presentó asociaciones relevantes con los temas 6, 2 "Estudio | Almidón de yuca" y 0 "Física | Órbitas celestes | Energía", a diferencia de las categorías "Engineering" y "Gasification | Tar | Downdraft gasifier" las cuales se consideraron insuficientes para comprender el tema del documento.

Gracias al tipo de implementación del método LDA realizada en este trabajo, es posible actualizar el modelo a medida que crezca el conjunto de datos, es decir, se realicen más publicaciones de artículos científicos o que se adapte a otros proyectos y tipos de conjuntos de datos. Igualmente, se puede extender la metodología a futuras incorporaciones de colecciones de artículos científicos de otras instituciones o uniendo estas colecciones para un análisis más profundo explotando el poder de estos modelos cuando los conjuntos de datos crecen.

Un parte importante en el modelado de temas son las visualizaciones de los temas latentes que permiten ayudar en la búsqueda, exploración y análisis documental, así como también facilitan ver relaciones complejas y complementarias entre temas para algunos artículos científicos de trabajos de investigación interdisciplinarios o innovadores, identificar temas de investigación emergentes y rezagados, determinar temas consolidados y de alto impacto, entre otros.

Un conjunto de datos de artículos científicos en Inglés, da la oportunidad de realizar incorporaciones ya sea de más documentos en inglés o en otro idioma, como por ejemplo en Español. Para lo cual se proponen métodos como traducción a un único idioma o técnicas de asociación de términos en ambos idiomas (e.g. *word embedding*), entre otros. Todo esto abre la puerta para trabajos futuros con el conjunto de datos consolidado en este trabajo.

Este tipo de trabajos resultan más fáciles para evaluar cuantitativamente por el tipo de resultados y técnicas existentes, ya que por la parte cualitativa, la gran cantidad de variables, o dimensiones, a evaluar complican la aplicación de técnicas comúnmente conocidas. Sin embargo, las evaluaciones cualitativas son un buen medio para comparar, analizar, interpretar y soportar las evaluaciones y resultados cuantitativos.

7.2. Recomendaciones y trabajo futuro

Un aprendizaje importante derivado del presente trabajo es el reconocimiento de la importancia de conocer el contexto y realizar la apropiada exploración del conjunto de datos con que se está trabajando dado que esto permite preconcebir e interpretar resultados parciales y finales. Igualmente, cada modelo LDA, o cualquier otro método computacional para tareas relacionadas, debe ser cuidadosamente estudiado y parametrizado dado que estos modelos son muy dependientes del conjunto y de la naturaleza de los datos.

Como trabajo futuro se propone extender el conjunto de datos ya sea con más artículos en Inglés o incorporando documentos en otro idioma como por ejemplo en Español, dado que en la región muchas publicaciones también se realizan en dicho idioma, por ejemplo, la Universidad de los Llanos con 113 al momento de consolidar el conjunto de datos. Además de extender el proceso metodológico a otras instituciones de educación superior de la región. También se propone extender el alcance de este trabajo a otros problemas del modelado de temas que permitirían obtener mayor información como por ejemplo la evolución temporal y la jerarquía de temas. Adicionalmente, incluir más elementos de los documentos para otros tipos de análisis y asociaciones como relaciones de autores, de citas o incluso mixtos.

De igual forma, se propone incorporar investigadores evaluadores adicionales tanto internos como externos a la institución y hacer el análisis sobre una muestra aleatoria, o por conveniencia más pequeña, al conjunto de datos total que sea representativa para un análisis más detallado y profundo.

Por último, se requiere más trabajo en la visualización de temas, dado que se logró identificar algunas relaciones entre temas y documentos, pero que, aún pueden mejorarse y trabajar en profundidad en el contexto de artículos científicos para contribuir al análisis documental,

gestión del conocimiento, adopción tecnológica, entre otros.

Bibliografía

[Acevedo, 2011] Acevedo, M. H. (2011). El Proceso De Codificación En Investigación Cualitativa. *Contribuciones a las Ciencias Sociales*, (2011-05).

[Aigneren, 1999] Aigneren, M. (1999). Análisis De Contenido. Una Introducción. *La Sociología en sus Escenarios*, 0(3).

[Alghamdi and Alfalqi, 2015] Alghamdi, R. and Alfalqi, K. (2015). A Survey of Topic Modeling in Text Mining. *International Journal of Advanced Computer Science and Applications*, 6(1):147–153.

[Allahyari et al., 2017] Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., and Kochut, K. (2017). A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques. *Proceedings of KDD Bigdas*, pages 1–13.

[Alpaydin, 2014] Alpaydin, E. (2014). *Introduction to Machine Learning, Third Edition*. The MIT Press, 3 edition.

[Andreu and Sieber, 1999] Andreu, R. and Sieber, S. (1999). La gestion integral del conocimiento y del aprendizaje. *Economia Industrial*, (326):63–72.

[Arrivillaga et al., 2016] Arrivillaga, J., Greenleaf, D., Hawthorn, M., and Alvarado, R. (2016). Revealing the landscape: Detecting trends in a scientific corpus. In *2016 IEEE Systems and Information Engineering Design Symposium (SIEDS)*, pages 292–297. IEEE.

[Banco Mundial, 2021] Banco Mundial (2021). PIB (UMN a precios actuales) - Colombia.

[Barreto, 2019] Barreto, L. M. (2019). Estudio de dos paradigmas de modelado de tópicos en un corpus de documentos tomados de una red social. Technical report, Universidad Central de Venezuela, Caracas, Venezuela.

[Bayes, 1763] Bayes, T. (1763). LII. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F. R. S. communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S. *Philosophical Transactions of the Royal Society of London*, 53:370–418.

[Bhardwaj and Khosla, 2017] Bhardwaj, P. and Khosla, P. (2017). Review of Text Mining Techniques. *IITM Journal of Management and IT*, 8(1):27–31.

[Bird et al., 2009] Bird, S., Klein, E., and Loper, E. (2009). *Natural language processing with Python*. O'Reilly Media; 1 edition (July 10, 2009), 1 edition.

[Blei, 2012] Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4):77.

[Blei and Lafferty, 2004] Blei, D. M. and Lafferty, J. D. (2004). Correlated Topic Models. In *Advances in Neural Information Processing Systems 18*, Vancouver.

[Blei et al., 2003] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022.

[Boyd-Graber et al., 2017] Boyd-Graber, J., Hu, Y., and Mimno, D. (2017). *Applications of Topic Models*, volume 11. Now Foundations and Trends.

[Cano et al., 2018] Cano, F., Cruz-Roa, A. A., and Madabhushi, A. (2018). A comparative analysis of sensitivity of convolutional neural networks for histopathology image classification in breast cancer. In Romero, E., Lepore, N., and Brieva, J., editors, *14th International Symposium on Medical Information Processing and Analysis*, volume 10975, page 39. SPIE.

[Chen et al., 2019] Chen, H., Wang, X., Pan, S., and Xiong, F. (2019). Identify Topic Relations in Scientific Literature Using Topic Modeling. *IEEE Transactions on Engineering Management*, pages 1–13.

[COLCIENCIAS, 2015] COLCIENCIAS (2015). Modelo de medición de grupos de investigación, desarrollo tecnológico o de innovación y de reconocimiento de investigadores. Technical report, Departamento Administrativo de Ciencia, Tecnología e Innovación - Colciencias, Bogotá.

[COLCIENCIAS, 2018a] COLCIENCIAS (2018a). Centros/Institutos de Investigación.

[COLCIENCIAS, 2018b] COLCIENCIAS (2018b). Comparativo Investigadores. Technical report, Departamento Administrativo de Ciencia, Tecnología e Innovación - Colciencias, Bogotá D.C.

[Congreso de Colombia, 2018a] Congreso de Colombia (2018a). Ley 1940 de 2018.

[Congreso de Colombia, 2018b] Congreso de Colombia (2018b). Ley 1942 de 2018.

[CONPES, 2014] CONPES (2014). Política para el desarrollo integral de la Orinoquia: Altillanura - Fase 1. (Documento CONPES 3797). Technical report, Consejo Nacional de Política Económica y Social - CONPES, Bogotá D.C.

[CONPES, 2018] CONPES (2018). Estrategia para la implementación de los objetivos de desarrollo sostenible (ODS) en Colombia. Technical report, Consejo Nacional de Política Económica y Social - CONPES, Bogotá D.C.

[CPC, 2019a] CPC (2019a). Índice Departamental de Competitividad 2019. Technical report, Consejo Privado de Competitividad (CPC), Bogotá D.C.

[CPC, 2019b] CPC (2019b). Informe Nacional De Competitividad 2019-2020. Technical report, Consejo Privado de Competitividad (CPC), Bogotá D.C.

[CPC, 2020] CPC (2020). Informe Nacional De Competitividad 2020-2021. Technical report, Consejo Privado de Competitividad (CPC), Bogotá D.C.

[De Battisti et al., 2015] De Battisti, F., Ferrara, A., and Salini, S. (2015). A decade of research in statistics: a topic model approach. *Scientometrics*, 103(2):413–433.

[Departamento Nacional de Planeación, 2017] Departamento Nacional de Planeación (2017). Sistema General de Regalías SGR. Technical report, Departamento Nacional de Planeación, Bogotá D.C.

[DNP, 2020] DNP (2020). Índice Departamental de Innovación para Colombia (IDIC) 2019. Technical report, Departamento Nacional de Planeación (DNP), Bogotá DC.

[Dutton and Conroy, 1997] Dutton, D. M. and Conroy, G. V. (1997). A review of machine learning. *The Knowledge Engineering Review*, 12(4):341–367.

[Estévez-Bretón, 2010] Estévez-Bretón, J. B. (2010). El desarrollo económico de la Orinoquia como aprendizaje y construcción de instituciones. *Colombia 2010-2014: Propuestas De Política Pública*, page 375–420.

[Feldman et al., 1998] Feldman, R., Fresko, M., Hirsh, H., Aumann, Y., Liphstat, O., Schler, Y., and Rajman, M. (1998). Knowledge Management: A Text Mining Approach. *Proc of the 2nd Int Conf on Practical Aspects of Knowledge Management - PAKM98*, Basel, Swi(April 2016):1–10.

[García Clausó, 1994] García Clausó, A. (1994). Fundamentos científicos del análisis documental. *Revista General de Información y Documentación*, 4(1):79.

[Gavel and Iselid, 2008] Gavel, Y. and Iselid, L. (2008). Web of Science and Scopus: A journal title overlap study. *Online Information Review*, 32(1):8–21.

[González Gil and Cano Arana, 2010] González Gil, T. and Cano Arana, A. (2010). Introducción al análisis de datos en investigación cualitativa:: Tipos de análisis y proceso de codificación (II). *NURE investigación: Revista Científica de enfermería, ISSN-e 1697-218X, Nº. 45, 2010*, (45):9.

[González Jiménez, 2001] González Jiménez, F. E. (2001). Generación del conocimiento y actividad educativa. *Revista Complutense de Educación*, 12(2):427–484.

[Gudivada et al., 2018] Gudivada, V. N., Rao, D. L., and Gudivada, A. R. (2018). Information Retrieval: Concepts, Models, and Systems. In *Handbook of Statistics*, volume 38, pages 331–401. Elsevier B.V.

[Guerrero-Casas, 2012] Guerrero-Casas, F. M. (2012). El análisis de escalamiento multidimensional: Una alternativa y un complemento a otras técnicas multivariantes. *La Sociología en sus Escenarios*, 0(25 SE - Metodología de la investigación social).

[Hamilton Wilson and Pezo Paredes, 2005] Hamilton Wilson, M. and Pezo Paredes, A. (2005). *Instrumentos de gestion de la ciencia, la tecnologia y la innovacion*. Convenio Andres Bello.

[Hanani et al., 2001] Hanani, U., Shapira, B., and Shoval, P. (2001). Information Filtering: Overview of Issues, Research and Systems. *User Modeling and User-Adapted Interaction*, 11(3):203–259.

[Heidenreich, 2018] Heidenreich, H. (2018). Stemming? Lemmatization? What? - Towards Data Science.

[Hernández-Sampieri et al., 2014] Hernández-Sampieri, R., Fernández-Collado, C., and Baptista-Lucio, P. (2014). *Metodología de la Investigación*. McGraw-Hill, México D.F., 6ªedición edition.

[Hoffman et al., 2010] Hoffman, M. D., Blei, D. M., and Bach, F. (2010). Online Learning for Latent Dirichlet Allocation. In *Advances in Neural Information Processing Systems 23 (NIPS 2010)*, pages 856–864.

[Hofmann, 1999] Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 1999*, pages 50–57. Association for Computing Machinery, Inc.

[Hofmann, 2001] Hofmann, T. (2001). Unsupervised learning by probabilistic Latent Semantic Analysis. *Machine Learning*, 42(1-2):177–196.

[Interactive Chaos, 2020] Interactive Chaos (2020). t-SNE — Interactive Chaos.

[Jelodar et al., 2019] Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., and Zhao, L. (2019). Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey. *Multimedia Tools and Applications*, 78(11):15169–15211.

[Jiménez Noblejas and Perianes Rodríguez, 2014] Jiménez Noblejas, C. and Perianes Rodríguez, A. (2014). Recuperación y visualización de información en Web of Science y Scopus: una aproximación práctica. *Investigación bibliotecológica*, 28(64).

[Kong et al., 2017] Kong, X., Jiang, H., Wang, W., Bekele, T. M., Xu, Z., and Wang, M. (2017). Exploring dynamic research interest and academic influence for scientific collaborator recommendation. *Scientometrics*, 113(1):369–385.

[Kurdi, 2016] Kurdi, M. Z. (2016). *Natural Language Processing and Computational Linguistics 1*. John Wiley & Sons, Inc., Hoboken, NJ, USA, 1 edition.

[Mandelbrot, 1953] Mandelbrot, B. (1953). An informational theory of the statistical structure of language. In Willis, J., editor, *Communication theory: Papers Read at a Symposium on "Applications of Communication Theory"*, pages 486–502, London. Butterworths.

[Manning et al., 2008] Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

[Martín-Martín et al., 2018] Martín-Martín, A., Orduna-Malea, E., Thelwall, M., and López-Cózar, E. D. (2018). Google Scholar, Web of Science, and Scopus: a systematic comparison of citations in 252 subject categories. *Journal of Informetrics*, 12(4):1160–1177.

[Miguel, 2011] Miguel, S. (2011). Revistas y producción científica de América Latina y el Caribe: su visibilidad en SciELO, RedALyC y SCOPUS. *Revista Interamericana de Bibliotecología*, 34(2).

[Ministerio de Educación de Colombia - Mineducación, 2010] Ministerio de Educación de Colombia - Mineducación (2010). Instituciones de Educación Superior.

[OCyT, 2020] OCyT (2020). Indicadores CTI 2019. Technical report, Observatorio Colombiano de Ciencia y Tecnología (OCyT).

[OEC, 2018] OEC (2018). Colombia (COL) Exports, Imports, and Trade Partners — OEC - The Observatory of Economic Complexity.

[Osorio Núñez, 2003] Osorio Núñez, M. (2003). El capital intelectual en la gestión del conocimiento. In *ACIMED*, volume 11, pages 0–0. Centro Nacional de Informacion de Ciencias Medicas, La habana.

[Pedregosa et al., 2011] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine Learning in {P}ython. *Journal of Machine Learning Research*, 12:2825–2830.

[Piantadosi, 2014] Piantadosi, S. T. (2014). Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin and Review*, 21(5):1112–1130.

[Pinto-Prieto et al., 2012] Pinto-Prieto, L.-P., Becerra-Ardila, L.-E., and Gómez-Flórez, L.-C. (2012). Carencias en los sistemas de gestión del conocimiento: una revisión bibliográfica. *El Profesional de la Informacion*, 21(3):268–276.

[Porras-García et al., 2018] Porras-García, Y. F., Calderon-Moreno, R., and Cruz-Roa, A. (2018). Análisis de Desempeño Computacional del Procesamiento Distribuido de una Implementación de Bolsa de Palabras en Apache Spark TM. In *2018 IEEE Colombian Conference on Communications and Computing (COLCOM)*, pages 1–6, Medellín. IEEE.

[Porter, 1980] Porter, M. (1980). An algorithm for suffix stripping. *Program*, 14(3):130–137.

[Rehurek and Sojka, 2010] Rehurek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA.

[Röder et al., 2015] Röder, M., Both, A., and Hinneburg, A. (2015). Exploring the space of topic coherence measures. *WSDM 2015 - Proceedings of the 8th ACM International Conference on Web Search and Data Mining*, pages 399–408.

[Sathya and Rajendran, 2013] Sathya, S. and Rajendran, N. (2013). A Review on Text Mining Techniques. *International Journal of Computer Science Trends and Technology (IJCST)*, 3(5):274–284.

[Shiryaev et al., 2017] Shiryaev, A. P., Dorofeev, A. V., Fedorov, A. R., Gagarina, L. G., and Zaycev, V. V. (2017). LDA models for finding trends in technical knowledge domain. In *2017 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus)*, pages 551–554. IEEE.

[Sievert and Shirley, 2014] Sievert, C. and Shirley, K. (2014). LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, pages 63–70, Baltimore. Association for Computational Linguistics (ACL).

[Soler-Tovar and Hernández-Rodríguez, 2018] Soler-Tovar, D. and Hernández-Rodríguez, P. (2018). Desarrollos y perspectivas de investigación en la Orinoquía. *Revista de Medicina Veterinaria*, (36):7–13.

[Srinivasa-Desikan, 2018] Srinivasa-Desikan, B. (2018). *Natural Language Processing and Computational Linguistics: A practical guide to text analysis with Python, Gensim, spaCy, and Keras*. Packt Publishing Limited, United Kingdom, Birmingham.

[Syed and Spruit, 2017] Syed, S. and Spruit, M. (2017). Full-Text or Abstract? Examining Topic Coherence Scores Using Latent Dirichlet Allocation. In *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 165–174. IEEE.

[Syed and Spruit, 2018] Syed, S. and Spruit, M. (2018). Selecting Priors for Latent Dirichlet Allocation. In *2018 IEEE 12th International Conference on Semantic Computing (ICSC)*, pages 194–202. IEEE.

[Torres-López and Arco-García, 2016] Torres-López, C. and Arco-García, L. (2016). Representación textual en espacios vectoriales semánticos. *Revista Cubana de Ciencias Informáticas*, 10(2):148–180.

[U-Sapiens, 2020] U-Sapiens (2020). Ranking de las mejores universidades de Colombia 2020 — U-Sapiens.

[van der Maaten and Hinton, 2008] van der Maaten, L. and Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605.

[Voorhees, 1999] Voorhees, E. M. (1999). Natural Language Processing and Information Retrieval. In *Lecture Notes in Computer Science - LNCS*, pages 32–48. Springer, Berlin, Heidelberg, vol 1714 edition.

[Xiong et al., 2019] Xiong, H., Cheng, Y., Zhao, W., and Liu, J. (2019). Analyzing scientific research topics in manufacturing field using a topic model. *Computers & Industrial Engineering*, 135:333–347.

[Yau et al., 2014] Yau, C.-K., Porter, A., Newman, N., and Suominen, A. (2014). Clustering scientific documents with topic modeling. *Scientometrics*, 100(3):767–786.

[Yoon and Park, 2004] Yoon, B. and Park, Y. (2004). A text-mining-based patent network: Analytical tool for high-technology trend. *The Journal of High Technology Management Research*, 15(1):37–50.

[Zhou et al., 2017] Zhou, H.-k., Yu, H.-m., and Hu, R. (2017). Topic discovery and evolution in scientific literature based on content and citations. *Frontiers of Information Technology & Electronic Engineering*, 18(10):1511–1524.

[Zipf, 1949] Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Addison-Wesley, Cambridge, MA.

A. Anexo: Listado 137 artículos del conjunto de datos en Inglés

El conjunto de datos base referente a documentos en Inglés, está disponible en Kaggle[1]. A continuación, en la tabla **A-2**, se muestran cuatro (4) de los 56 metadatos del listado de los 137 artículos científicos que conforman este conjunto de datos.

Tabla A-2.: Listado 137 artículos del conjunto de datos en Inglés. Cuatro (4) de 56 metadatos.

ID	Autor	Título	DOI
0	Mendoza D.L.S., Baquero J.E.M., Varela O.M.A.	Monitoring system of relativy humidity, CO, CO2, NH3 and temperature control for small shed	10.18687/LACCEI2019.1.1.271
1	Niño W.D.P., Jiménez F.R.L., Jiménez A.F.L.	Algorithms to estimation of size and shape tomato using Artificial Vision Techniques	10.18687/LACCEI2019.1.1.5
2	Vargas J., Varela Ó.A., Valera Á.	Numerical and analytical analysis of a 3UPS-2RPRRR parallel robot	10.18687/LACCEI2019.1.1.307
3	CorredorSantamaría W., Torres-Tabares A., VelascoSantamaría Y.M.	Biochemical and histological alterations in Aequidens metae (Pisces, Cichlidae) and Astyanax gr. bimaculatus (Pisces, Characidae) as indicators of river pollution	10.1016/j.scitotenv.2019.07.187
4	SalinasJimenez L.G., Boldrini R., OsorioRamirez D.P., Caro C.I., RojasPeña J.I.	A new species of Camelobaetidius Demoulin, 1966 (Ephemeroptera: Baetidae), from the Colombian Orinoco River basin	10.11646/zootaxa.4656.2.9
5	Zhang J., Wang X., Zhu Y., Huang Z., Yu Z., Bai Y., Fan G., Wang P., Chen H., Su Y., TrujilloGonzález J.M., Hu B.X., Krebs P., Hua P.	The influence of heavy metals in road dust on the surface runoff quality: Kinetic, isotherm, and sequential extraction investigations	10.1016/j.ecoenv.2019.03.106
6	Parra A.S., MoraDelgado J.	Emission factors estimated from enteric methane of dairy cattle in Andean zone using the IPCC Tier2 methodology	10.1007/s1045701701773
7	CalderónDelgado I.C., MoraSolarte D.A., VelascoSantamaría Y.M.	Physiological and enzymatic responses of Chlorella vulgaris exposed to produced water and its potential for bioremediation	10.1007/s1066101975198
8	TrujilloGonzález J.M., TorresMora M.A., JiménezBallesta R., Zhang J.	Landusedependent spatial variation and exposure risk of heavy metals in roaddeposited sediment in Villavicencio, Colombia	10.1007/s1065301801606
9	Rodrigues D.M., Hunter L.G., Bernard F.L., Rojas M.F., Dalla Vecchia F., Einloft S.	Harnessing CO 2 into Carbonates Using Heterogeneous Waste Derivative Cellulose-Based Poly(ionic liquids) as Catalysts	10.1007/s1056201826374
10	Campbell S., Bernard F.L., Rodrigues D.M., Rojas M.F., Carreño L., Chaban V.V., Einloft S.	Performance of metalfunctionalized rice husk cellulose for CO2 sorption and CO2/N2 separation	10.1016/j.fuel.2018.11.078

Continúa en la siguiente página

[1]Conjunto de datos disponible: DS_Unillanos_Papers_EN

Tabla A-2 – continuación de la página anterior

ID	Autor	Título	DOI
11	Ellis L.T., Afonina O.M., Doroshina G.Y., Agudelo C., Andriamiarisoa R.L., Asthana A.K., Gupta D., Gupta R., Rawat K.K., Sahu V., Aymerich P., BednarekOchyra H., Brugués M., Ruiz E., Sáez L., Callaghan D.A., Caspari S., Drapela P., Dugarova O.D., Tubanova D.Y., Erzberger P., Flores J.R., Suárez G.M., Fedosov V.E., Gospodinov G., Gradstein S.R., Reeb C., Jukonienė I., Subkaitė M., Kučera J., Lee G.E., Lombo Y.J., Suarez K.Y., Lebouvier M., Majumdar S., Müller F., Nagy J., Norhazrina N., Papp B., Plášek V., Pócs T., Puglisi M., SchäferVerwimp A., Shirzadian S., Singh D.K., Ştefănuţ S., Torzewski K., van Melick H., Wolski G.J., Zander R.H.	New national and regional bryophyte records, 58	10.1080/03736687.2018.1559636
12	SalamancaRamos E., Velasco Páez Z.J., Baquero Álvarez N.	Health status of the elderly in life centers [Estado de saúde dos idosos dos centros de vida] [Estado de salud de los adultos mayores de los centros vida]	10.5294/aqui.2019.19.2.3
13	Díaz D., Corredor G., Romero E., CruzRoa A.	A webbased telepathology framework for collaborative work of pathologists to support teaching and research in latin america	10.1007/9783030138356_12
14	Dubeibe F.L., MartínezSicachá S.M., González G.A.	Orbital dynamics in realistic galaxy models: NGC 3726, NGC 3877 and NGC 4010	10.18257/raccefyn.774
15	Jaimes D.A.R., Contreras D., Jimenez A.M.F., OrcioliSilva D., Barbieri F.A., Gobbi L.T.B.	Effects of linear and undulating periodization of strength training in the acceleration of skater children	10.1590/s1980-6574201900010007
16	Portacio A.A., Rodríguez B.A., Villamil P.	Theoretical study on optical response in nanostructures in the Born–Markov regime: The role of spontaneous emission and dephasing	10.1016/j.aop.2018.11.023
17	Hernández G., Jimenez A.F., Ortiz B.V., Lamadrid A.P., Cardenas P.F.	Decision Support System for Precision Irrigation Using Interactive Maps and Multiagent Concepts	10.1007/9783030044473_2
18	Burbano R.P., Carrascal A.K., Arango J.L.P., Bautista J.L.R.	Assessment of a multiplex detection method for Salmonella enterica, Escherichia coli O157:H7, and Listeria monocytogenes in cow milk	10.11144/JAVERIANA.SC24-1.AOAM
19	RojasMolina Y.A., Provenzano-Rizzi F., RamírezGil H.	A new species of whiptail armored catfish, genus Pseudohemiodon (Siluriformes: Loricariidae) from the Orinoco river basin, Llanos region of Colombia and Venezuela	10.1590/1982022420180160
20	Coelho G.C.Z., Yo I.S., MiraLópez T.M., Monzani P.S., Arashiro D.R., Fujimoto T., Senhorini J.A., Yasui G.S.	Preparing a fish embryo (Prochilodus lineatus) for staging, chorion removal and PGC traceability	10.1387/ijdb.180348gc
21	Jimenez A.F., Herrera E.F., Ortiz B.V., Ruiz A., Cardenas P.F.	Inference System for Irrigation Scheduling with an Intelligent Agent	10.1007/9783030044473_1
22	Zotos E.E., Dubeibe F.L., Nagler J., Tejeda E.	Orbit classification in a pseudoNewtonian Copenhagen problem with Schwarzschildlike primaries	10.1093/mnras/stz1432
23	Vargas J.H., Varela O.A., Valera A.	Geometric Analysis of a 3R2T Low Mobility Parallel Robot	10.1109/CCRA.2018.8588154
24	BernalPáez C., Sánchez F.	Harvest rates and foraging strategy of Carollia perspicillata (Chiroptera: Phyllostomidae) in an artificial food patch	10.1016/j.beproc.2018.07.010
25	LondoñoBurbano A., Urbano-Bonilla A., RojasMolina Y., RamírezGil H., PradaPedreros S.	A New Species of Spatuloricaria Schultz, 1944 (Siluriformes: Loricariidae), from the Orinoco River Basin, Colombia	10.1643/CI18087
26	Barros de Freitas A.C., Ortiz Vega W.H., Quirino C.R., Bartholazzi Junior A., Gomes David C.M., Geraldo A.T., Silva Rua M.A., Cipagauta Rojas L.F., Eustáquio de Almeida Filho J., Burla Dias A.J.	Surface temperature of ewes during estrous cycle measured by infrared thermography	10.1016/j.theriogenology.2018.07.015
27	Freitas A.C.B.D., Quirino C.R., Bartholazzi Junior A., Vega W.H.O., David C.M.G., Geraldo A.T., Rua M.A.S., Rojas L.F.C., Almeida Filho J.E.D., Dias A.J.B.	Surface temperature in different anatomical regions of ewes measured by infrared thermography	10.1016/j.livsci.2018.07.014
28	Cardona C.C.C., Coronado Y.M., Coronado A.C.M., Ochoa I.	Genetic diversity in oil palm (Elaeis guineensis jacq) using RAM (random amplified microsatellites)	10.1590/16784499.2017385

Continúa en la siguiente página

Tabla A-2 – continuación de la página anterior

ID	Autor	Título	DOI
29	Romero G.A., Guzman C., Rueda L., Castro J.R., Cardenas A., Agbossou K.	Case Study of Data Management for Power and Energy Monitoring	10.1109/ISIE.2018.8433731
30	JiménezOjeda Y.K., CollazosLasso L.F., AriasCastellanos J.A.	Dynamics and use of nitrogen in biofloc technology BFT	
31	OchoaAmaya J.E., QueirozHazarbassanov N., Namazu L.B., Calefi A.S., Tobaruela C.N., Margatho R., PalermoNeto J., Ligeiro De Oliveira A.P., Felicio L.F.	ShortTerm Hyperprolactinemia Reduces the Expression of Purinergic P2X7 Receptors during Allergic Inflammatory Response of the Lungs	10.1159/000489312
32	Millan L.M.R., Dominguez M.A.C., Vargas F.E.S.	Effect of temperature and air equivalence ratio on energy potential of syngas produced from oil palm shells gasification	10.15866/ireme.v12i7.14379
33	Avendaño J.E., TejeiroM. N., DíazCárdenas J., AmayaBurgos J.J., Aponte A.F., Gamboa N., José SalcedoSarmiento Y.E., VelásquezSuárez Á.J., MoralesRozo A.	Birds of universidad de los Llanos (Villavicencio, Colombia): A rich community at the andean foothillssavanna transition [Aves de la universidad de los Llanos (Villavicencio, Colombia): Una rica comunidad en la transición entre el piedemonte andino y la sabana]	10.17151/bccm.2018.22.2.5
34	Zotos E.E., Dubeibe F.L., González G.A.	Orbit classification in an equalmass nonspinning binary black hole pseudoNewtonian system	10.1093/MNRAS/STY946
35	Rua M.A.S., Quirino C.R., Ribeiro R.B., Carvalho E.C.Q., Bernadino M.D.L.A., Bartholazzi Junior A., Cipagalta L.F., Barreto M.A.P.	Diagnostic methods to detect uterus illnesses in mares	10.1016/j.theriogenology.2018.03.042
36	Grenfell G.G., Nascimento I.C., Oliveira D.S., GuimarãesFilho Z.O., Elizondo J.I., Reis A.P., Galvão R.M.O., Baquero W.A.H., Oliveira A.M., Ronchi G., De Sá W.P., Severo J.H.F.	Hmode access and the role of spectral shift with electrode biasing in the TCABR tokamak	10.1063/1.5029561
37	CollazosLasso L.F., GutiérrezEspinosa M.C., AyaBaquero E.	Induced reproduction of the sailfin pleco, pterygoplichthys gibbiceps (Kner, 1854) (pisces: Loricariidae)	
38	CruzRoa A., Gilmore H., Basavanhally A., Feldman M., Ganesan S., Shih N., Tomaszewski J., Madabhushi A., González F.	Highthroughput adaptive sampling for wholeslide histopathology image analysis (HASHI) via convolutional neural networks: Application to invasive breast cancer detection	10.1371/journal.pone.0196828
39	Dubeibe F.L., RiañoDoncel A., Zotos E.E.	Dynamical analysis of bounded and unbounded orbits in a generalized Hénon–Heiles system	10.1016/j.physleta.2018.02.001
40	Pereira F.A.C., Hernandez W.A., Toufen D.L., GuimarãesFilho Z.O., Caldas I.L., Gentle K.W.	Burst temperature from conditional analysis in Texas Helimak and TCABR tokamak	10.1063/1.5025062
41	Zotos E.E., RiañoDoncel A., Dubeibe F.L.	Basins of convergence of equilibrium points in the generalized Hénon–Heiles system	10.1016/j.ijnonlinmec.2017.12.004
42	Zotos E.E., Dubeibe F.L.	Orbital dynamics in the postNewtonian planar circular restricted SunJupiter system	10.1142/S0218271818500360
43	MurilloPacheco J., LópezIborra G.M., Escobar F., BonillaRojas W.F., Verdú J.R.	The value of small, natural and manmade wetlands for bird diversity in the east Colombian Piedmont	10.1002/aqc.2835
44	Lozano E., Calderón R., Enciso J.	A multistrategy recommendation algorithm to retrieve broken hyperlinks	10.18687/LACCEI2018.1.1.140
45	SalinasJiménez L.G., RojasPeña J.I., OsorioRamírez D.P., CaroCaro C.I.	Erratum: New records of ephemeroptera from the colombian orinoco river basin of the meta department (Revista Colombiana de Entomología, (2017) 43, 2 (271276), 10.25100/socolen.v43i2.5958)	10.25100/socolen.v44i1.6758
46	RamirezNiño M.Á., JiménezForero J.A., BernalSalazar J.P., OsorioDueñas M.D.	Characterization of oil extracted from the kernel of the fruit of cumare's palm (Astrocaryum chambira barret) [Caracterización del aceite extraído del kernel del fruto de la palma de cumare (Astrocaryum chambira Barret)]	10.15446/rfna.v71n1.69589
47	Cano F., Madabhushi A., CruzRoa A.	A comparative analysis of sensitivity of convolutional neural networks for histopathology image classification in breast cancer	10.1117/12.2511647
48	Zapata D., CruzRoa A., Jiménez A.	Automatic classification of optical defects of mirrors from ronchigram images using bag of visual words and support vector machines	10.1007/9783319751931_86
49	Portacio A.A., Rodríguez B.A., Villamil P.	Influence of the position of a donor impurity on the secondorder nonlinear optical susceptibility in a cylindrical quantum dot	10.1016/j.spmi.2017.11.041

Continúa en la siguiente página

Tabla A-2 – continuación de la página anterior

ID	Autor	Título	DOI
50	Herrera W.D., Leon O.L., Londono W.C., Vargas J.A.	Implementation of a control and biometric safety of the vascular network of the dorsal side of the hand through digital processing of images	10.1109/CCAC.2017.8276408
51	MoralesRozo A., Tenorio E.A., Carling M.D., Cadena C.D.	Origin and crosscentury dynamics of an avian hybrid zone	10.1186/s1286201710967
52	Morales M.M.B., Hernández J.J.M., PetiniBenelli A.	A new species of Catasetum (Orchidaceae: Catasetinae) from Casanare, Colombia	10.15517/lank.v17i3.31644
53	Ramirez J.L., Birindelli J.L., Carvalho D.C., Affonso P.R.A.M., Venere P.C., Ortega H., CarrilloAvila M., RodríguezPulido J.A., Galetti P.M.	Revealing hidden diversity of the underestimated neotropical ichthyofauna: DNA barcoding in the recently described genus Megaleporinus (characiformes: Anostomidae)	10.3389/fgene.2017.00149
54	HoyosLeyva J.D., AlonsoGomez L., RuedaEnciso J., YeeMadeira H., BelloPerez L.A., AlvarezRamirez J.	Morphological, physicochemical and functional characteristics of starch from Marantha ruiziana Koern	10.1016/j.lwt.2017.05.019
55	Riahi Manesh M., Subramaniam S., Reyes H., Kaabouch N.	Realtime spectrum occupancy monitoring using a probabilistic model	10.1016/j.comnet.2017.06.003
56	LópezBarragan C.N., Sánchez F.	Food selection and predation risk in the Andean whiteeared opossum (Didelphis pernigra Allen, 1900) in a suburban area of Bogotá, Colombia	10.1016/j.mambio.2017.07.001
57	SalinasJiménez L.G., RojasPeña J.I., OsorioRamírez D.P., CaroCaro C.I.	New records of ephemeroptera from the colombian orinoco river basin of the meta department [Nuevos registros de ephemeroptera para la cuenca colombiana del orinoco en el departamento de meta]	10.25100/socolen.v43i2.5958
58	TrujilloGonzález J.M., MahechaPulido J.D., TorresMora M.A., Brevik E.C., Keesstra S.D., JiménezBallesta R.	Impact of potentially contaminated river water on agricultural irrigated soils in an equatorial climate	10.3390/agriculture7070052
59	Elderini T., Kaabouch N., Reyes H.	Channel quality estimation metrics in cognitive radio networks: A survey	10.1049/ietcom.2016.0919
60	Riahi Manesh M., Kaabouch N., Reyes H., Hu W.C.	A Bayesian approach to estimate and model SINR in wireless networks	10.1002/dac.3187
61	JaimesDueñez J., TrianaChávez O., CantilloBarraza O., Hernández C., Ramírez J.D., GóngoraOrjuela A.	Molecular and serological detection of Trypanosoma cruzi in dogs (Canis lupus familiaris) suggests potential transmission risk in areas of recent acute Chagas disease outbreaks in Colombia	10.1016/j.prevetmed.2017.03.009
62	Dubeibe F.L., LoraClavijo F.D., González G.A.	On the conservation of the Jacobi integral in the postNewtonian circular restricted threebody problem	10.1007/s1050901730761
63	CruzRoa A., Gilmore H., Basavanhally A., Feldman M., Ganesan S., Shih N.N.C., Tomaszewski J., González F.A., Madabhushi A.	Accurate and reproducible invasive breast cancer detection in wholeslide images: A Deep Learning approach for quantifying tumor extent	10.1038/srep46450
64	Portacio A.A., Rodríguez B.A., Villamil P.	Nonlinear optical response of an impurity in a cylindrical quantum dot under the action of a magnetic field	10.1016/j.physb.2017.02.008
65	Elderini T., Kaabouch N., Reyes H.	Outage probability estimation technique based on a Bayesian model for cognitive radio networks	10.1109/CCWC.2017.7868355
66	Dubeibe F.L., LoraClavijo F.D., González G.A.	PseudoNewtonian planar circular restricted 3body problem	10.1016/j.physleta.2016.12.024
67	VelascoSantamaría Y.M., TorresTabares A., RamírezSaray J.A., CruzCasallas P.E., RamírezMerlano J.A., QuirogaSanchez É., AyaBaquero E.	Feeding habits of leporinus friderici (Anostomidae: Teleostei) during a hydrobiological cycle in vaupés river, Colombia [Hábitos alimenticios de Leporinus friderici (Anostomidae: Teleostei) durante un ciclo hidrobiológico en el río Vaupés, Colombia]	10.15517/rbt.v65i2.22929
68	Vega O., Duarte H., Chavarriaga J.	Software development process supported by business process modeling an experience report	
69	BarreraRojas L.M., ObandoBastidas J.A., PuelloMendez J.	Technological profile analysis in dairy companies: A case study	10.3303/CET1757296
70	Franco M.C., Domenech M., Costa J.L., Aparicio V.	Modelling effective soil depth at field scale from soil sensors and geomorphometric indices	10.15446/acag.v66n2.53282
71	Riaño J., Paqui M.F., CórdobaCórdoba S., Sánchez F.	Nest and chicks of Pseudoscops clamator (Aves: Strigidae) in the highland plateau of the sabana de Bogotá, Colombia [Nido y polluelos de Pseudoscops clamator (Aves: strigidae) en el altiplano de la sabana de Bogotá, Colombia]	10.15446/abc.v22n1.54380
72	Reyes H., Subramaniam S., Kaabouch N., Hu W.C.	A Bayesian inference method for estimating the channel occupancy	10.1109/UEMCON.2016.7777865

Continúa en la siguiente página

Tabla A-2 – continuación de la página anterior

ID	Autor	Título	DOI
73	Manesh M.R., Kaabouch N., Reyes H., Hu W.C.	A Bayesian model of the aggregate interference power in cognitive radio networks	10.1109/UEMCON.2016.7777828
74	CorredorSantamaría W., Serrano Gómez M., VelascoSantamaría Y.M.	Using genotoxic and haematological biomarkers as an evidence of environmental contamination in the Ocoa River native fish, Villavicencio—Meta, Colombia	10.1186/s4006401617530
75	AlonsoGomez L., NiñoLópez A.M., RomeroGarzón A.M., PinedaGomez P., del RealLopez A., Rodriguez-Garcia M.E.	Physicochemical transformation of cassava starch during fermentation for production of sour starch in Colombia	10.1002/star.201600059
76	Martinez N., Rodriguez Martinez J.A.	Firstprinciples study on the formation energies of Ga1xCrxAs	10.1088/1742-6596/743/1/012007
77	Dubeibe F.L., SanabriaGómez J.D.	Geodesic motion in a stationary dihole spacetime	10.1103/PhysRevD.94.044058
78	Jimenez Lopez A.F., Prieto Pelayo M.C., Ramirez Forero A.	Teaching Image Processing in Engineering Using Python	10.1109/RITA.2016.2589479
79	Avendaño J.E., Barker F.K., Cadena C.D.	The Yellowgreen Bushtanager is neither a bushtanager nor a sparrow: Molecular phylogenetics reveals that Chlorospingus flavovirens is a tanager (Aves: Passeriformes; Thraupidae)	10.11646/zootaxa.4136.2.7
80	PaQui M.F., MuñozGaray J., MantillaMeluk H., Sánchez F.	First record of Promops nasutus (Spix, 1823) (Chiroptera: Molossidae) from Colombia	10.15560/12.3.1915
81	Morales M.M.B., AguirreMorales C., Cardenas J.	Passiflora creucicaetanoae a new species of Passiflora L. supersection Tacsonia (Passifloraceae) from Colombia	10.11646/phytotaxa.261.3.6
82	TrujilloGonzález J.M., TorresMora M.A., Keesstra S., Brevik E.C., JiménezBallesta R.	Heavy metal accumulation related to population density in road dust samples taken from urban sites under different land uses	10.1016/j.scitotenv.2016.02.101
83	Reyes H., Subramaniam S., Kaabouch N., Hu W.C.	A spectrum sensing technique based on autocorrelation and Euclidean distance and its comparison with energy detection for cognitive radio networks	10.1016/j.compeleceng.2015.05.015
84	MurilloPacheco J.I., BonillaRojas W.F.	New records and distribution extensions of some bird species in the Colombian Andean-Orinoco, department of Meta	10.15560/12.2.1876
85	Rios J., Romero C.A., Molina D.	Instrumentation and control of a DC motor through a web platform	10.1109/IESummit.2016.7459776
86	OchoaAmaya J.E., Marino L.P., Tobaruela C.N., Namazu L.B., Calefi A.S., Margatho R., Gonçalves V., Jr., QueirozHazarbassanov N., Klein M.O., PalermoNeto J., De Oliveira A.P.L., Cristina De O.M., Felicio L.F.	Attenuated allergic inflammatory response in the lungs during lactation	10.1016/j.lfs.2016.03.027
87	Colmenares P. C.H., Silva P. A., Mogollón O. Á.M.	Impacts of different coffee systems on soil microbial populations at different altitudes in Villavicencio (Colombia) [Impactos de diferentes sistemas de café sobre las poblaciones microbiales del suelo a diferentes altitudes en Villavicencio (Colombia)]	10.15446/agron.colomb.v34n2.55420
88	RamírezGil H.	Spatial and temporal length distribution of Zungaro zungaro caught in the Orinoco River Basin of Colombia	
89	Sanabria M.L.V., de Rodríguez L.M.	Needs of parents in caring for their children in a pediatric intensive care unit	10.17533/udea.iee.v34n1a04
90	Criollo E. H., Silva P. A., Delgado H. H.	Greenhouse gas balance related to conventional and sustainable fruit production systems in the highlands region of Pasto, Colombia [Balance de gases de efecto invernadero relacionado a sistemas convencionales y sostenibles de producción de frutas en la región del Altiplano de Pasto, Colombia]	10.15446/agron.colomb.v34n2.55417
91	OcampoPeñuela N., PeñuelaRecio L., OcampoDurán Á.	Decals prevent birdwindow collisions at residences: A successful case study from Colombia [Calcomanías evitan colisiones de aves contra ventanas de residencias: Estudio de un caso exitoso de Colombia]	
92	RamosMolina L.M., ChavarroMesa E., Pereira D.A.S., SilvaHerrera M.R., Ceresini P.C.	Rhizoctonia solani AG1 IA infects both rice and signalgrass in the Colombian Llanos [Rhizoctonia solani AG1 IA infecta arroz e braquiária nos Llanos Colombianos]	10.1590/1983-40632016v4638696
93	MurilloPacheco J.I., Rös M., Escobar F., CastroLima F., Verdú J.R., LópezIborra G.M.	Effect of wetland management: Are lentic wetlands refuges of plantspecies diversity in the AndeanOrinoco Piedmont of Colombia?	10.7717/peerj.2267
94	Rios J., Romero C.A., Molina D.	Instrumentation and control of a DC motor through the Ubidots platform	10.1109/WEA.2015.7370121

Continúa en la siguiente página

Tabla A-2 – continuación de la página anterior

ID	Autor	Título	DOI
95	OchoaAmaya J.E., Hamasato E.K., Tobaruela C.N., Queiroz-Hazarbassanov N., Anselmo Franci J.A., PalermoNeto J., Greiffo F.R., De Britto A.A., Vieira R.P., Ligeiro De Oliveira A.P., Massoco C.D.O., Felicio L.F.	Shortterm hyperprolactinemia decreases allergic inflammatory response of the lungs	10.1016/j.lfs.2015.10.016
96	RuizGarcía M., PinedoCastro M., LuengasVillamil K., Vergara C., Rodriguez J.A., Shostell J.M.	Molecular phylogenetics of the whitelipped peccary (Tayassu pecari) did not confirm morphological subspecies in northwestern South America	10.4238/2015.May.22.6
97	Avendaño J.E., Cuervo A.M., LópezO. J.P., GutiérrezPinto N., CortésDiago A., Cadena C.D.	A new species of tapaculo (Rhinocryptidae: Scytalopus) from the Serranía de Perijá of Colombia and Venezuela	10.1642/AUK14166.1
98	Dittrich T., Dubeibe F.L.	Classical and quantum chaotic angular-momentum pumps	10.1103/PhysRevLett.114.094101
99	Gómez M.C.O., Gómez M.C.O.	Cultural tourism in Villavicencio Colombia	10.1007/9783319057354_6
100	Subramaniam S., Reyes H., Kaabouch N.	Spectrum occupancy measurement: An autocorrelation based scanning technique using USRP	10.1109/WAMICON.2015.7120376
101	Avendaño J.E., Donegan T.M.	A distinctive new subspecies of scytalopus griseicollis (Aves, passeriformes, rhinocryptidae) from the northern eastern cordillera of Colombia and Venezuela	10.3897/zookeys.506.9553
102	Donegan T.M., Avendaño J.E.	'Bogotá' type specimens of the hummingbird genus Adelomyia, with diagnosis of an overlooked subspecies from the East Andes of Colombia	
103	CruzRoa A., Arévalo J., Judkins A., Madabhushi A., González F.	A method for medulloblastoma tumor differentiation based on convolutional neural networks and transfer learning	10.1117/12.2208825
104	LópezO. J.P., Avendaño J.E., GutiérrezPinto N., Cuervo A.M.	The birds of the Serranía de Perijá: The northernmost avifauna of the Andes [Las aves de la Serranía de Perijá: La avifauna más septentrional de los Andes]	
105	LongmanMills S., Williams Y.M.G., Rodriguez M.O.M., Baquero M.R.G., Rojas J.D.G., de Amaya C.J., Diaz E.A.M., Corea S.J.P., Baez E.M.P., Tinoco L.I.S.	The association between adult drug abuse and childhood maltreatment in students attending seven universities in five countries in Latin America and one country in the Caribbean [A associacao entre o abuso de drogas e o maltrato infantil em estudantes de sete universidades de cinco paises da América Latina e um do Caribe] [La asociacion entre el abuso de drogas y el maltrato infantil en estudiantes de siete uniersidades de cinco paises de Latin America y uno pais del Caribe]	10.1590/0104-07072015001ESP026
106	Rao I., Ishitani M., Miles J., Peters M., Tohme J., Arango J., Moreta D.E., Lopez H., Castro A., Van Der Hoek R., Martens S., Hyman G., Tapasco J., Duitama J., Suárez H., Borrero G., Núñez J., Hartmann K., Domínguez M., Sotelo M., Vergara D., Lavelle P., Subbarao G.V., Rincon A., Plazas C., Mendoza R., Rathjen L., Karwat H., Cadisch G.	Climatesmart croplivestock systems for smallholders in the tropics: Integration of new forage hybrids to intensify agriculture and to mitigate climate change through regulation of nitrification in soil	10.17138/TGFT(2)130132
107	Dubeibe F.L., BermúdezAlmanza L.D.	Optimal conditions for the numerical calculation of the largest Lyapunov exponent for systems of ordinary differential equations	10.1142/S0129183114500247
108	Gucrrcro H.B., Baquero Velasquez A.E., Barrero J.F., Cöco D.Z., Risardi J.C., Magalhães D.V., Becker M.	Orientation (Yaw) Fuzzy controller applied to a carlike mobile robot prototype	10.1109/CWCAS.2014.6994603
109	Morillo C. Y., Morillo C. A.C., Muñoz F. J.E., Ballesteros P. W., González A.	Molecular characterization of 93 genotypes of cocoa (Theobroma cacao L.) with random amplified microsatellites RAMs [Caracterización molecular con microsatélites amplificados al azar (RAMs) de 93 genotipos de cacao (Theobroma cacao L.)]	10.15446/agron.colomb.v32n3.46879
110	Donegan T.M., Avendaño J.E., Lambert F.	A new Tapaculo related to Scytalopus rodriguezi from Serranía de los Yariguíes, Colombia	
111	Livengood E.J., Aya E., Arias J.A., Chapman F.A.	Quantitative measurement of epithelial injury in ornamental silver dollar fish (Metynnis orinocensis) captured in the wild, imported wildcaught, and aquacultured	

Continúa en la siguiente página

Tabla A-2 – continuación de la página anterior

ID	Autor	Título	DOI
112	ChavarroRodríguez N., Díaz-Castelazo C., RicoGray V.	Characterization and functional ecology of the extrafloral nectar of Cedrela odorata in contrasting growth environments in central Veracruz, Mexico	10.1139/cjb20120289
113	VelascoSantamaría Y.M., Bjerregaard P., Korsgaard B.	Evidence of small modulation of ethinylestradiol induced effects by concurrent exposure to trenbolone in male eelpout Zoarces viviparus	10.1016/j.envpol.2013.03.011
114	Trujillo A.V., Reina A.E.G., Orjuela A.G., Suárez E.P., Palomares J.E., Alvarez L.S.B.	Seasonal variation and natural infection of lutzomyia antunesi (Diptera: Psychodidae: Phlebotominae), an endemic species in the orinoquia region of Colombia	10.1590/S0074-0276108042013011
115	LongmanMills S., González W.Y., Meléndez M.O., García M.R., Gómez J.D., Juárez C.G., Martínez E.A., Peñalba S.J., Pizzanelli E.M., Solórzano L.I., Wright M.G.M., Cumsille F., De La Haye W., Sapag J.C., Khenti A., Hamilton H.A., Erickson P.G., Brands B., Flam-Zalcman R., Simpson S., Wekerle C., Mann R.E.	Exploring child maltreatment and its relationship to alcohol and cannabis use in selected Latin American and Caribbean countries	10.1016/j.chiabu.2012.11.002
116	Borchsenius F., Surez L.S.S., Prince L.M.	Molecular phylogeny and redefined generic limits of Calathea (Marantaceae)	10.1600/036364412X648571
117	RamírezDuarte W.F., Pineda-Quiroga C., Martínez N., Eslava-Mocha P.R.	Use of sodium chloride and zeolite during shipment of Ancistrus triradiatus under high temperature	10.1590/S1679-62252011005000036
118	VelascoSantamaría Y.M., Korsgaard B., Madsen S.S., Bjerregaard P.	Bezafibrate, a lipidlowering pharmaceutical, as a potential endocrine disruptor in male zebrafish (Danio rerio)	10.1016/j.aquatox.2011.05.018
119	OchoaAmaya J.E., Malucelli B.E., CruzCasallas P.E., Nasello A.G., Felicio L.F., CarvalhoFreitas M.I.R.	Dual effects of hyperprolactinemia on carrageenaninduced inflammatory paw edema in rats	10.1159/000323774
120	RamirezMerlano J.A., Velasco-Santamaría Y.M., MedinaRobles V.M., CruzCasallas P.E.	Cryopreservation effects on the sperm quality of cachama blanca Piaractus brachypomus (Cuvier 1818)	10.1111/j.1365-2109.2011.02835.x
121	CruzCasallas P.E., MedinaRobles V.M., VelascoSantamaría Y.M.	Fish farming of native species in Colombia: Current situation and perspectives	10.1111/j.1365-2109.2011.02855.x
122	VelascoSantamaría Y.M., Handy R.D., Sloman K.A.	Endosulfan affects health variables in adult zebrafish (Danio rerio) and induces alterations in larvae development	10.1016/j.cbpc.2011.01.001
123	Pachón L.A., Dubeibe F.L.	The influence of the Lande gfactor in the classical general relativistic description of atomic and subatomic systems	10.1088/0264-9381/28/5/055002
124	ChavesBedoya G., Espejel F., AlcaláBriseo R.I., HernándezVela J., SilvaRosales L.	Short distance movement of genomic negative strands in a host and nonhost for Sugarcane mosaic virus (SCMV)	10.1186/1743422X815
125	VelascoSantamaría Y., Corredor-Santamaría W.	Nutritional requirements of freshwater ornamental fish: A review	
126	LongmanMills S., González Y.W., Meléndez M.O., García M.R., Gómez J.D., Juárez C.G., Martínez E.A., Peñalba S.J., Pizzanelli M.E., Solórzano L.I., Wright G.M., Cumsille F., Sapag J.C., Wekerle C., Hamilton H.A., Erickson P.G., Mann R.E.	Child Maltreatment and Its Relationship to Drug Use in Latin America and the Caribbean: An Overview and Multinational Research Partnership	10.1007/s1146901193470
127	SanabriaGómez J.D., Hernández-Pastora J.L., Dubeibe F.L.	Innermost stable circular orbits around magnetized rotating massive stars	10.1103/PhysRevD.82.124014
128	Dubeibe F.L.	Solving the timedependent schrödinger equation with absorbing boundary conditions and source terms in mathematica 6.0	10.1142/S0129183110015919
129	OchoaAmaya J.E., Malucelli B.E., CruzCasallas P.E., Nasello A.G., Felicio L.F., CarvalhoFreitas M.I.R.	Acute and chronic stress and the inflammatory response in hyperprolactinemic rats	10.1159/000292063
130	VelascoSantamaría Y.M., Cruz-Casallas P.E.	Behavioural and gill histopathological effects of acute exposure to sodium chloride in moneda (Metynnis orinocensis)	10.1016/j.etap.2007.12.002
131	CruzCasallas P.E., MedinaRobles V.M., VelascoSantamaría Y.M.	Seasonal variation of sperm quality and the relationship between spermatocrit and sperm concentration in yamú Brycon amazonicus	10.1577/A06002.1
132	VelascoSantamaría Y.M., Medina-Robles V.M., CruzCasallas P.E.	Cryopreservation of yamú (Brycon amazonicus) sperm for large scale fertilization	10.1016/j.aquaculture.2006.02.039
133	CruzCasallas P.E., Lombo-Rodríguez D.A., VelascoSantamaría Y.M.	Milt quality and spermatozoa morphology of captive Brycon siebenthalae (Eigenmann) broodstock	10.1111/j.1365-2109.2005.01273.x

Continúa en la siguiente página

Tabla A-2 – continuación de la página anterior

ID	Autor	Título	DOI
134	Holmann F., Rivas L., Urbina N., Rivera B., Giraldo L.A., Guzman S., Martinez M., Medina A., Ramirez G.	The role of livestock in poverty alleviation: An analysis of Colombia	
135	Hucke E.E.T.S., CruzCasallas P.E., Sider L.H., Felicio L.F.	Reproductive experience modulates dopaminerelated behavioral responses	10.1016/S00913057(01)004580
136	CruzCasallas P.E., Felicio L.F., Nasello A.G.	A quantitative analysis of the role of experience in the regulation of sexual behavior in male rats	10.3758/BF03331998

B. Anexo: Listado de temas dominantes por documento

Tabla B-2.: Listado de temas dominante por cada documento

ID Doc.	Tema dom.	Prob. Tema	Términos Top-10 del tema	Título del documento	Térms. Top-10 del Doc. según Tema Dom.
0	6	0.772	result, study, sample, different, increase, model, value, effect, use, data	"*Monitoring system of relativy humidity, CO, CO2, NH3 and temperature control for small shed*"	['control', 'design', 'response', 'phase', 'temperature', 'sensor', 'source_author', 'level', 'environment', 'necessary']
1	6	0.729	result, study, sample, different, increase, model, value, effect, use, data	"*Algorithms to estimation of size and shape tomato using Artificial Vision Techniques*"	['image', 'obtain', 'size', 'implement', 'application', 'process', 'calculate', 'algorithm', 'classification', 'define']
2	0	0.996	case, value, orbit, order, correspond, present, point, energy, time, potential	"*Numerical and analytical analysis of a 3UPS-2RPRRR parallel robot*"	['point', 'analysis', 'solution', 'value', 'base', 'equations', 'platform', 'cos', 'vector', 'sin']
3	1	0.661	fish, sh, observe, exposure, water, effect, concentration, endosulfan, gill, liver	"*Biochemical and histological alterations in Aequidens metae (Pisces, Cichlidae) and Astyanax gr. bimaculatus (Pisces, Characidae) as indicators of river pollution*"	['site', 'rainy_season', 'sh', 'metae', 'liver', 'oxidative_stress', 'gr_bimaculatus', 'ocoa_river', 'gill', 'water']
4	8	0.64	species, colombia, en, study, del, specimens, bird, record, include, base	"*A new species of Camelobaetidius Demoulin, 1966 (Ephemeroptera: Baetidae), from the Colombian Orinoco River basin*"	['length', 'new_species', 'zootaxa', 'round', 'setae', 'colombia', 'apex', 'species', 'thomas', 'zootaxa_magnolia_press']
5	9	0.65	soil, metal, heavy_metal, sediment, urban, mg, water, source, adsorption, wastewater	"*The influence of heavy metals in road dust on the surface runoff quality: Kinetic, isotherm, and sequential extraction investigations*"	['adsorption', 'rd', 'metal', 'heavy_metal', 'equilibrium', 'surface', 'mg', 'isotherm', 'initial', 'soil']
6	4	0.703	cattle, ipcc, livestock, soil, emissions, farm, systems, pasture, variables, emission_factor	"*Emission factors estimated from enteric methane of dairy cattle in Andean zone using the IPCC Tier-2 methodology*"	['ipcc', 'emission_factor', 'ch', 'ym', 'methane', 'animal', 'prp', 'pasture', 'milk', 'cattle']
7	6	0.467	result, study, sample, different, increase, model, value, effect, use, data	"*Physiological and enzymatic responses of Chlorella vulgaris exposed to produced water and its potential for bioremediation*"	['control', 'different', 'increase', 'treatment', 'high', 'data', 'study', 'table', 'report', 'measure']
8	9	0.763	soil, metal, heavy_metal, sediment, urban, mg, water, source, adsorption, wastewater	"*Land-use-dependent spatial variation and exposure risk of heavy metals in road-deposited sediment in Villavicencio, Colombia*"	['metal', 'exposure', 'mg', 'land_use', 'soil', 'zhang', 'pollution', 'risk', 'pli', 'heavy_metal']
9	2	0.891	starch, cellulose, peak, increase, sample, properties, compound, cassava_starch, nitrogen, higher	"*Harnessing CO 2 into Carbonates Using Heterogeneous Waste Derivative Cellulose-Based Poly(ionic liquids) as Catalysts*"	['catalyst', 'cpils', 'catalysts', 'cpil_tbp', 'increase', 'catalytic', 'catalytic_activity', 'yield', 'cpil', 'pressure']
10	2	0.698	starch, cellulose, peak, increase, sample, properties, compound, cassava_starch, nitrogen, higher	"*Performance of metal-functionalized rice husk cellulose for CO2 sorption and CO2/N2 separation*"	['cellulose', 'tio', 'fe', 'cellulose_tio', 'sorption', 'surface', 'peak', 'adsorption', 'magnetite', 'ti']
11	8	0.717	species, colombia, en, study, del, specimens, bird, record, include, base	"*New national and regional bryophyte records, 58*"	['species', 'south', 'eastern', 'record', 'report', 'ochyra', 'bednarek_ochyra', 'forest', 'leg', 'leave']
12	6	0.511	result, study, sample, different, increase, model, value, effect, use, data	"*Health status of the elderly in life centers [Estado de saúde dos idosos dos centros de vida] [Estado de salud de los adultos mayores de los centros vida]*"	['study', 'health', 'age', 'care', 'condition', 'lc', 'functional', 'women', 'years_age', 'dependence']

Continúa en la siguiente página

Tabla B-2 – continuación de la página anterior

ID Doc.	Tema dom.	Prob. Tema	Términos Top-10 del tema	Título del documento	Términos Top-10 del Doc. según el Tema Dominante
13	6	0.948	result, study, sample, different, increase, model, value, effect, use, data	"*A web-based telepathology framework for collaborative work of pathologists to support teaching and research in latin america*"	['navigation', 'magni_cation', 'pathologists', 'slide_histopathology_image', 'web_base_telepathology_framework', 'time', 'telepathology', 'di_erent', 'image', 'concurrent_users']
14	0	0.901	case, value, orbit, order, correspond, present, point, energy, time, potential	"*Orbital dynamics in realistic galaxy models: NGC 3726, NGC 3877 and NGC 4010*"	['model', 'galaxies', 'galaxy_model', 'orbit', 'analytical', 'miyamoto_nagai', 'rotation_curve', 'energy', 'lz', 'star']
15	6	0.9	result, study, sample, different, increase, model, value, effect, use, data	"*Effects of linear and undulating periodization of strength training in the acceleration of skater children*"	['train', 'exercise', 'level', 'effect', 'strength_train', 'performance', 'acceleration', 'group', 'speed', 'sport']
16	0	0.844	case, value, orbit, order, correspond, present, point, energy, time, potential	"*Theoretical study on optical response in nanostructures in the Born–Markov regime: The role of spontaneous emission and dephasing*"	['quantum', 'optical', 'eq', 'term', 'quantum_dot', 'expressions', 'eqs', 'optical_field', 'time', 'phenomenological']
17	6	0.973	result, study, sample, different, increase, model, value, effect, use, data	"*Decision Support System for Precision Irrigation Using Interactive Maps and Multi-agent Concepts*"	['crop', 'water', 'information', 'irrigation', 'model', 'eld', 'use', 'need', 'elds', 'agents']
18	6	0.925	result, study, sample, different, increase, model, value, effect, use, data	"*Assessment of a multiplex detection method for Salmonella enterica, Escherichia coli O157:H7, and Listeria monocytogenes in cow milk*"	['food', 'strain', 'coli', 'listeria_monocytogenes', 'milk', 'monocytogenes', 'ufc_ml', 'pcr', 'sel_broth', 'detection']
19	8	0.952	species, colombia, en, study, del, specimens, bird, record, include, base	"*A new species of whiptail armored catfish, genus Pseudohemiodon (Siluriformes: Loricariidae) from the Orinoco river basin, Llanos region of Colombia and Venezuela*"	['species', 'pseudohemiodon', 'river', 'plat', 'isbrücker_nijssen', 'length', 'head', 'isbrücker', 'iavh_mm_sl', 'ray']
20	6	0.917	result, study, sample, different, increase, model, value, effect, use, data	"*Preparing a fish embryo (Prochilodus lineatus) for staging, chorion removal and PGC traceability*"	['stage', 'embryos', 'chorion', 'hatch', 'pgcs', 'development', 'egg', 'stag', 'embryo', 'fish']
21	6	0.974	result, study, sample, different, increase, model, value, effect, use, data	"*Inference System for Irrigation Scheduling with an Intelligent Agent*"	['irrigation', 'water', 'control', 'soil_moisture', 'temperature', 'base', 'soil', 'crop', 'model', 'agent']
22	0	0.942	case, value, orbit, order, correspond, present, point, energy, time, potential	"*Orbit classification in a pseudo-Newtonian Copenhagen problem with Schwarzschild-like primaries*"	['orbit', 'case', 'rs', 'correspond', 'pn', 'close_encounter', 'motion', 'area', 'primaries', 'potentials']
23	0	0.926	case, value, orbit, order, correspond, present, point, energy, time, potential	"*Geometric Analysis of a 3R-2T Low Mobility Parallel Robot*"	['point', 'platform', 'joint', 'parallel_robots', 'rb', 'base', 'parallel_robot', 'limbs', 'rotation', 'online_available']
24	6	0.937	result, study, sample, different, increase, model, value, effect, use, data	"*Harvest rates and foraging strategy of Carollia perspicillata (Chiroptera: Phyllostomidae) in an artificial food patch*"	['bat', 'food', 'information', 'feeders', 'forage', 'patch', 'time', 'fruit', 'use', 'obtain']
25	8	0.985	species, colombia, en, study, del, specimens, bird, record, include, base	"*A New Species of Spatuloricaria Schultz, 1944 (Siluriformes: Loricariidae), from the Orinoco River Basin, Colombia*"	['spatuloricaria', 'river', 'plat', 'colombia', 'length', 'vs', 'spatuloricaria_terracanticum', 'include', 'river_basin', 'species']
26	6	0.988	result, study, sample, different, increase, model, value, effect, use, data	"*Surface temperature of ewes during estrous cycle measured by infrared thermography*"	['temperature', 'phase', 'estrus', 'ewes', 'accept_manuscript', 'ovulation', 'irt', 'infrared_thermography', 'period', 'temperatures']
27	6	0.986	result, study, sample, different, increase, model, value, effect, use, data	"*Surface temperature in different anatomical regions of ewes measured by infrared thermography*"	['temperature', 'irt', 'estrous_cycle', 'accept_manuscript', 'surface_temperature', 'vulva', 'ewes', 'ovulation', 'infrared_thermography', 'ear']
28	6	0.938	result, study, sample, different, increase, model, value, effect, use, data	"*Genetic diversity in oil palm (Elaeis guineensis jacq) using RAM (random amplified microsatellites)*"	['oil_palm', 'genetic_diversity', 'genotypes', 'evaluate', 'germplasm', 'markers', 'value', 'ram', 'populations', 'oil_palm_elaeis_guineensis']

Continúa en la siguiente página

Tabla B-2 – continuación de la página anterior

ID Doc.	Tema dom.	Prob. Tema	Términos Top-10 del tema	Título del documento	Términos Top-10 del Doc. según el Tema Dominante
29	6	0.989	result, study, sample, different, increase, model, value, effect, use, data	"*Case Study of Data Management for Power and Energy Monitoring*"	['data', 'data_management', 'database', 'smart', 'management', 'systems', 'storage', 'power', 'energy', 'measurement']
30	6	0.518	result, study, sample, different, increase, model, value, effect, use, data	"*Dynamics and use of nitrogen in biofloc technology - BFT*"	['nitrite', 'culture', 'different', 'water', 'aquaculture', 'report', 'aacl_bioflux_volume_issue', 'bioflux_com_ro_aacl', 'systems', 'generate']
31	6	0.983	result, study, sample, different, increase, model, value, effect, use, data	"*Short-Term Hyperprolactinemia Reduces the Expression of Purinergic P2X7 Receptors during Allergic Inflammatory Response of the Lungs*"	['prolactin', 'expression', 'receptor', 'cells', 'prlr', 'rat', 'lung', 'asthma', 'induce', 'increase']
32	6	0.454	result, study, sample, different, increase, model, value, effect, use, data	"*Effect of temperature and air equivalence ratio on energy potential of syngas produced from oil palm shells gasification*"	['value', 'oil_palm', 'energetic', 'right_reserve', 'obtain', 'er', 'range', 'agent', 'acceptable', 'copyright']
33	8	0.84	species, colombia, en, study, del, specimens, bird, record, include, base	"*Birds of universidad de los Llanos (Villavicencio, Colombia): A rich community at the andean foothills-savanna transition [Aves de la universidad de los Llanos (Villavicencio, Colombia): Una rica comunidad en la transición entre el piedemonte andino y la sabana]*"	['species', 'bird', 'link', 'forest', 'campus', 'colombia', 'avifauna', 'habitats', 'record', 'species_richness']
34	0	0.972	case, value, orbit, order, correspond, present, point, energy, time, potential	"*Orbit classification in an equal-mass non-spinning binary black hole pseudo-Newtonian system*"	['case', 'orbit', 'primaries', 'rs', 'download_academic_oup_com', 'mnras_advance_article_abstract', 'mnras_sty_university_durham', 'energy', 'plane', 'correspond']
35	6	0.985	result, study, sample, different, increase, model, value, effect, use, data	"*Diagnostic methods to detect uterus illnesses in mares*"	['mar', 'uterine', 'cytology', 'culture', 'endometritis', 'cytobrush', 'accept_manuscript', 'biopsy', 'endometrial', 'lvc']
36	0	0.696	case, value, orbit, order, correspond, present, point, energy, time, potential	"*H-mode access and the role of spectral shift with electrode biasing in the TCABR tokamak*"	['bias', 'plasma', 'turbulence_suppression', 'transition', 'shear', 'edge', 'shift', 'turbulence', 'radial', 'turbulent']
37	8	0.586	species, colombia, en, study, del, specimens, bird, record, include, base	"*Induced reproduction of the sailfin pleco, pterygoplichthys gibbiceps (Kner, 1854) (pisces: Loricariidae)*"	['species', 'colombia', 'sailfin_pleco', 'ehc', 'pp_spanish', 'aya_baquero', 'ornamental_fish', 'reproductive', 'females', 'ortega_lara']
38	6	0.982	result, study, sample, different, increase, model, value, effect, use, data	"*High-throughput adaptive sampling for whole-slide histopathology image analysis (HASHI) via convolutional neural networks: Application to invasive breast cancer detection*"	['sample', 'image', 'cnn', 'approach', 'hashi', 'wsi', 'regions', 'detection', 'tile', 'train']
39	0	0.951	case, value, orbit, order, correspond, present, point, energy, time, potential	"*Dynamical analysis of bounded and unbounded orbits in a generalized Hénon–Heiles system*"	['value', 'energy', 'basin_entropy', 'potential', 'case', 'escape', 'chaos', 'ghh', 'hénon_heiles', 'emin']
40	0	0.744	case, value, orbit, order, correspond, present, point, energy, time, potential	"*Burst temperature from conditional analysis in Texas Helimak and TCABR tokamak*"	['burst', 'probe', 'texas_helimak', 'machine', 'plasma', 'density', 'fit', 'tcabr', 'ion_saturation_current', 'average']
41	0	0.903	case, value, orbit, order, correspond, present, point, energy, time, potential	"*Basins of convergence of equilibrium points in the generalized Hénon–Heiles system*"	['equilibrium_point', 'case', 'newton_raphson_basins_convergence', 'basins', 'panel', 'initial_condition', 'basin_entropy', 'correspond', 'present', 'basins_convergence']
42	0	0.944	case, value, orbit, order, correspond, present, point, energy, time, potential	"*Orbital dynamics in the post-Newtonian planar circular restricted Sun-Jupiter system*"	['orbit', 'case', 'initial_condition', 'correspond', 'pn', 'jupiter', 'value_jacobi_constant', 'panel', 'motion', 'type']

Continúa en la siguiente página

Tabla B-2 – continuación de la página anterior

ID Doc.	Tema dom.	Prob. Tema	Términos Top-10 del tema	Título del documento	Términos Top-10 del Doc. según el Tema Dominante
43	8	0.719	species, colombia, en, study, del, specimens, bird, record, include, base	"*The value of small, natural and man-made wetlands for bird diversity in the east Colombian Piedmont*"	['wetlands', 'wetland', 'species', 'bird', 'murillo_pacheco', 'conservation', 'bird_diversity', 'landbirds', 'natural_wetlands', 'study']
44	6	0.958	result, study, sample, different, increase, model, value, effect, use, data	"*A multi-strategy recommendation algorithm to retrieve broken hyperlinks*"	['search', 'web', 'internet', 'archive', 'result', 'strategy', 'hyperlinks', 'break', 'request', 'python']
45	8	0.963	species, colombia, en, study, del, specimens, bird, record, include, base	"*Erratum: New records of ephemeroptera from the colombian orinoco river basin of the meta department (Revista Colombiana de Entomología, (2017) 43, 2 (271-276), 10.25100/socolen.v43i2.5958)*"	['caro', 'meta_department', 'colombiana', 'palabras_clave', 'colombian_orinoco_river_basin', 'revista_colombiana_entomología', 'salinas_jiménez', 'new_record_ephemeroptera', 'rojas_peña', 'ramírez']
46	6	0.645	result, study, sample, different, increase, model, value, effect, use, data	"*Characterization of oil extracted from the kernel of the fruit of cumare's palm (Astrocaryum chambira barret) [Caracterización del aceite extraído del kernel del fruto de la palma de cumare (Astrocaryum chambira Barret)]*"	['palm', 'oil_extract', 'fruit', 'fatty_acids', 'sample', 'percentage', 'seed', 'obtain', 'report', 'ml']
47	6	0.989	result, study, sample, different, increase, model, value, effect, use, data	"*A comparative analysis of sensitivity of convolutional neural networks for histopathology image classification in breast cancer*"	['train', 'image', 'cnns', 'idc_cnn', 'breast_cancer', 'case', 'performance', 'hup', 'cnn', 'different']
48	6	0.844	result, study, sample, different, increase, model, value, effect, use, data	"*Automatic classification of optical defects of mirrors from ronchigram images using bag of visual words and support vector machines*"	['ronchigram_image', 'image', 'class', 'svm', 'ronchigrams', 'mirror', 'ronchi_test', 'sample', 'sift', 'train']
49	0	0.855	case, value, orbit, order, correspond, present, point, energy, time, potential	"*Influence of the position of a donor impurity on the second-order nonlinear optical susceptibility in a cylindrical quantum dot*"	['magnetic_eld', 'impurity', 'increase', 'quantum_dot', 'position_impurity', 'optical_recti_cation', 'effect', 'cqd', 'energy', 'value']
50	6	0.929	result, study, sample, different, increase, model, value, effect, use, data	"*Implementation of a control and biometric safety of the vascular network of the dorsal side of the hand through digital processing of images*"	['image', 'security', 'process', 'information', 'infrared', 'application', 'user', 'systems', 'allow', 'hand']
51	6	0.59	result, study, sample, different, increase, model, value, effect, use, data	"*Origin and cross-century dynamics of an avian hybrid zone*"	['hybrid_zone', 'model', 'estimate', 'data', 'time', 'value', 'result', 'base', 'sample', 'different']
52	8	0.939	species, colombia, en, study, del, specimens, bird, record, include, base	"*A new species of Catasetum (Orchidaceae: Catasetinae) from Casanare, Colombia*"	['species', 'catasetum', 'colombia', 'mm', 'labellum', 'elliptic', 'lucisuareziae', 'bonilla', 'catasetum_lucisuareziae', 'bicolor']
53	8	0.805	species, colombia, en, study, del, specimens, bird, record, include, base	"*Revealing hidden diversity of the underestimated neotropical ichthyofauna: DNA barcoding in the recently described genus Megaleporinus (characiformes: Anostomidae)*"	['motus', 'species', 'obtusidens', 'paraná', 'megaleporinus', 'nominal_species', 'reinhardti', 'macrocephalus', 'trifasciatus', 'ku']
54	2	0.601	starch, cellulose, peak, increase, sample, properties, compound, cassava_starch, nitrogen, higher	"*Morphological, physicochemical and functional characteristics of starch from Marantha ruiziana Koern*"	['starch', 'mrk_starch', 'cassava_starch', 'properties', 'gelatinization', 'amylose_content', 'starch_granules', 'crystallinity', 'higher', 'mrk']
55	6	0.982	result, study, sample, different, increase, model, value, effect, use, data	"*Real-time spectrum occupancy monitoring using a probabilistic model*"	['channel', 'bayesian_inference', 'probability', 'channel_occupancy', 'occupancy', 'spectrum_sense', 'time', 'frequentist_inference', 'include', 'result']
56	6	0.747	result, study, sample, different, increase, model, value, effect, use, data	"*Food selection and predation risk in the Andean white-eared opossum (Didelphis pernigra Allen, 1900) in a suburban area of Bogotá, Colombia*"	['opossums', 'food', 'predator', 'station', 'gud', 'feeders', 'dog_urine', 'affect', 'kotler', 'preference']

Continúa en la siguiente página

Tabla B-2 – continuación de la página anterior

ID Doc.	Tema dom.	Prob. Tema	Términos Top-10 del tema	Título del documento	Términos Top-10 del Doc. según el Tema Dominante
57	8	0.939	species, colombia, en, study, del, specimens, bird, record, include, base	"*New records of ephemeroptera from the colombian orinoco river basin of the meta department [Nuevos registros de ephemeroptera para la cuenca colombiana del orinoco en el departamento de meta]*"	['ephemeroptera', 'colombia', 'vereda', 'domínguez', 'dias', 'species', 'salinas', 'lugo_ortiz_mccafferty', 'molineri', 'en']
58	9	0.474	soil, metal, heavy_metal, sediment, urban, mg, water, source, adsorption, wastewater	"*Impact of potentially contaminated river water on agricultural irrigated soils in an equatorial climate*"	['soil', 'wastewater', 'crossref', 'heavy_metal', 'water', 'source', 'agriculture', 'area', 'irrigate', 'agricultural']
59	6	0.983	result, study, sample, different, increase, model, value, effect, use, data	"*Channel quality estimation metrics in cognitive radio networks: A survey*"	['channel', 'pp', 'metrics', 'represent', 'interference', 'sinr', 'parameters', 'channel_quality_estimation', 'ber', 'outage_probability']
60	6	0.957	result, study, sample, different, increase, model, value, effect, use, data	"*A Bayesian approach to estimate and model SINR in wireless networks*"	['sinr', 'model', 'state', 'variables', 'path_loss', 'base', 'different', 'nod', 'awgn', 'interference']
61	6	0.782	result, study, sample, different, increase, model, value, effect, use, data	"*Molecular and serological detection of Trypanosoma cruzi in dogs (Canis lupus familiaris) suggests potential transmission risk in areas of recent acute Chagas disease outbreaks in Colombia*"	['dog', 'trypanosoma_cruzi', 'transmission', 'infection', 'chagas_disease', 'sample', 'page', 'animals', 'cruzi', 'positive']
62	0	0.93	case, value, orbit, order, correspond, present, point, energy, time, potential	"*On the conservation of the Jacobi integral in the post-Newtonian circular restricted three-body problem*"	['post_newtonian', 'order', 'equations_motion', 'pn', 'jacobi_constant', 'newtonian', 'lagrangian', 'chaotic', 'derive', 'hamiltonian']
63	6	0.982	result, study, sample, different, increase, model, value, effect, use, data	"*Accurate and reproducible invasive breast cancer detection in whole-slide images: A Deep Learning approach for quantifying tumor extent*"	['case', 'train', 'slide', 'breast_cancer', 'image', 'slide_image', 'cohort', 'pathology', 'classifiers', 'feature']
64	0	0.925	case, value, orbit, order, correspond, present, point, energy, time, potential	"*Non-linear optical response of an impurity in a cylindrical quantum dot under the action of a magnetic field*"	['impurity', 'bind_energy', 'magnetic_eld', 'cqd', 'energy', 'quantum_dot', 'quantum', 'ect', 'optical_recti_cation', 'state']
65	6	0.966	result, study, sample, different, increase, model, value, effect, use, data	"*Outage probability estimation technique based on a Bayesian model for cognitive radio networks*"	['sinr', 'outage_probability', 'combination_prx_pi_pn', 'level', 'state', 'result', 'prx', 'channel', 'value', 'combination']
66	0	0.945	case, value, orbit, order, correspond, present, point, energy, time, potential	"*Pseudo-Newtonian planar circular restricted 3-body problem*"	['pseudo_newtonian', 'potential', 'orbit', 'crtbp', 'set', 'limit', 'source', 'study', 'set_initial_condition', 'dynamics']
67	8	0.98	species, colombia, en, study, del, specimens, bird, record, include, base	"*Feeding habits of leporinus friderici (Anostomidae: Teleostei) during a hydrobiological cycle in vaupés river, Colombia [Hábitos alimenticios de Leporinus friderici (Anostomidae: Teleostei) durante un ciclo hidrobiológico en el río Vaupés, Colombia]*"	['en', 'del', 'se', 'peces', 'para', 'por', 'especie', 'leporinus_friderici', 'mayor', 'cv']
68	6	0.967	result, study, sample, different, increase, model, value, effect, use, data	"*Software development process supported by business process modeling an experience report*"	['process', 'model', 'software', 'bpmn', 'business', 'rup', 'project', 'client', 'represent', 'stakeholders']
69	6	0.967	result, study, sample, different, increase, model, value, effect, use, data	"*Technological profile analysis in dairy companies: A case study*"	['company', 'technology', 'process', 'variables', 'technological', 'production', 'innovation', 'management', 'dairy', 'market']
70	6	0.879	result, study, sample, different, increase, model, value, effect, use, data	"*Modelling effective soil depth at field scale from soil sensors and geomorphometric indices*"	['esd', 'model', 'predictors', 'eca', 'field', 'rf', 'soil', 'effective_soil_depth', 'geomorphometric_indices', 'tosca']

Continúa en la siguiente página

Tabla B-2 – continuación de la página anterior

ID Doc.	Tema dom.	Prob. Tema	Términos Top-10 del tema	Título del documento	Términos Top-10 del Doc. según el Tema Dominante
71	8	0.84	species, colombia, en, study, del, specimens, bird, record, include, base	"*Nest and chicks of Pseudoscops clamator (Aves: Strigidae) in the highland plateau of the sabana de Bogotá, Colombia [Nido y polluelos de Pseudoscops clamator (Aves: strigidae) en el altiplano de la sabana de Bogotá, Colombia]*"	['nest', 'clamator', 'en', 'colombia', 'sánchez', 'pseudoscops_clamator', 'argentina', 'bogotá_highland_plateau', 'thurber', 'aves']
72	6	0.991	result, study, sample, different, increase, model, value, effect, use, data	"*A Bayesian inference method for estimating the channel occupancy*"	['bayesian_inference', 'sample', 'channel_occupancy_rate', 'channel', 'probability', 'probability_distribution', 'spectrum_decision', 'estimate', 'spectrum', 'process']
73	6	0.968	result, study, sample, different, increase, model, value, effect, use, data	"*A Bayesian model of the aggregate interference power in cognitive radio networks*"	['state', 'interference', 'model', 'path_loss', 'interference_power', 'variables', 'parent', 'frequency', 'aggregate_interference_power', 'variable']
74	1	0.652	fish, sh, observe, exposure, water, effect, concentration, endosulfan, gill, liver	"*Using genotoxic and haematological biomarkers as an evidence of environmental contamination in the Ocoa River native fish, Villavicencio—Meta, Colombia*"	['fish', 'observe', 'site', 'reference_site', 'water', 'erythrocytes', 'metae', 'season', 'genotoxic', 'rainy_season']
75	2	0.61	starch, cellulose, peak, increase, sample, properties, compound, cassava_starch, nitrogen, higher	"*Physicochemical transformation of cassava starch during fermentation for production of sour starch in Colombia*"	['starch', 'fermentation', 'peak', 'cassava_starch', 'sample', 'cassava', 'days', 'starch_granules', 'bacteria', 'process']
76	0	0.734	case, value, orbit, order, correspond, present, point, energy, time, potential	"*First-principles study on the formation energies of Ga1-xCrxAs*"	['gaas', 'structure', 'atom', 'formation_energy', 'ga_atom', 'cr', 'lattice', 'position', 'calculate', 'vacancy']
77	0	0.936	case, value, orbit, order, correspond, present, point, energy, time, potential	"*Geodesic motion in a stationary dihole spacetime*"	['orbit', 'photons', 'solution', 'case', 'dihole', 'parameters', 'massive_particles', 'orbit_coordinate', 'effective_potential', 'symmetry_axis']
78	6	0.951	result, study, sample, different, increase, model, value, effect, use, data	"*Teaching Image Processing in Engineering Using Python*"	['students', 'project', 'learn', 'image_process', 'develop', 'teach', 'image', 'software', 'development', 'course']
79	8	0.789	species, colombia, en, study, del, specimens, bird, record, include, base	"*The Yellow-green Bush-tanager is neither a bush-tanager nor a sparrow: Molecular phylogenetics reveals that Chlorospingus flavovirens is a tanager (Aves: Passeriformes; Thraupidae)*"	['flavovirens', 'species', 'tanagers', 'burn', 'genus', 'clade', 'bangsia', 'dx', 'thraupidae', 'analyse']
80	8	0.849	species, colombia, en, study, del, specimens, bird, record, include, base	"*First record of Promops nasutus (Spix, 1823) (Chiroptera: Molossidae) from Colombia*"	['promops', 'specimen', 'colombia', 'chiroptera', 'nasutus', 'molossidae', 'species', 'venezuela', 'iavh', 'specimens']
81	8	0.922	species, colombia, en, study, del, specimens, bird, record, include, base	"*Passiflora creuci-caetanoae a new species of Passiflora L. supersection Tacsonia (Passifloraceae) from Colombia*"	['col', 'municipality', 'passiflora', 'species', 'bonilla', 'leave', 'creuci_caetanoae', 'morales_jbb', 'colombia', 'vereda']
82	9	0.675	soil, metal, heavy_metal, sediment, urban, mg, water, source, adsorption, wastewater	"*Heavy metal accumulation related to population density in road dust samples taken from urban sites under different land uses*"	['soil', 'sediment', 'heavy_metal', 'urban', 'metal', 'dx', 'pollution', 'geo', 'water', 'brevik']
83	6	0.97	result, study, sample, different, increase, model, value, effect, use, data	"*A spectrum sensing technique based on autocorrelation and Euclidean distance and its comparison with energy detection for cognitive radio networks*"	['signal', 'snr', 'sample', 'usrp', 'noise', 'cognitive_radio', 'spectrum_sense', 'autocorrelation', 'different', 'receive']
84	8	0.922	species, colombia, en, study, del, specimens, bird, record, include, base	"*New records and distribution extensions of some bird species in the Colombian Andean-Orinoco, department of Meta*"	['species', 'colombia', 'bird', 'villavicencio', 'record', 'puerto_lópez', 'pdf', 'report', 'meta_department', 'pp']
85	6	0.949	result, study, sample, different, increase, model, value, effect, use, data	"*Instrumentation and control of a DC motor through a web platform*"	['control', 'devices', 'ubidots', 'data', 'time', 'process', 'design', 'speed', 'allow', 'controller']

Continúa en la siguiente página

Tabla B-2 – continuación de la página anterior

ID Doc.	Tema dom.	Prob. Tema	Términos Top-10 del tema	Título del documento	Términos Top-10 del Doc. según el Tema Dominante
86	6	0.982	result, study, sample, different, increase, model, value, effect, use, data	"*Attenuated allergic inflammatory response in the lungs during lactation*"	['lactation', 'asthma_lactation', 'group', 'prolactin', 'accept_manuscript_ochoa_amaya', 'asthma', 'immune', 'rat', 'increase', 'ifn']
87	6	0.761	result, study, sample, different, increase, model, value, effect, use, data	"*Impacts of different coffee systems on soil microbial populations at different altitudes in Villavicencio (Colombia) [Impactos de diferentes sistemas de café sobre las poblaciones microbiales del suelo a diferentes altitudes en Villavicencio (Colombia)]*"	['coffee', 'zone', 'bp', 'cfu', 'coffee_systems', 'factor', 'systems', 'fp', 'ap', 'different']
88	8	0.898	species, colombia, en, study, del, specimens, bird, record, include, base	"*Spatial and temporal length distribution of Zungaro zungaro caught in the Orinoco River Basin of Colombia*"	['en', 'se', 'cm', 'tmg', 'cm_ls', 'zungaro', 'females', 'fish', 'hembras', 'size']
89	6	0.858	result, study, sample, different, increase, model, value, effect, use, data	"*Needs of parents in caring for their children in a pediatric intensive care unit*"	['parent', 'care', 'nurse', 'children', 'need', 'information', 'research', 'theme', 'god', 'picu']
90	4	0.592	cattle, ipcc, livestock, soil, emissions, farm, systems, pasture, variables, emission_factor	"*Greenhouse gas balance related to conventional and sustainable fruit production systems in the highlands region of Pasto, Colombia [Balance de gases de efecto invernadero relacionado a sistemas convencionales y sostenibles de producción de frutas en la región del Altiplano de Pasto, Colombia]*"	['soil', 'variables', 'emissions', 'ipcc', 'systems', 'eq_ha_yr', 'crop_residues', 'total_ghg', 'potential', 'ghg_emissions']
91	8	0.61	species, colombia, en, study, del, specimens, bird, record, include, base	"*Decals prevent bird-window collisions at residences: A successful case study from Colombia [Calcomanías evitan colisiones de aves contra ventanas de residencias: Estudio de un caso exitoso de Colombia]*"	['bird', 'collisions', 'decals', 'windows', 'species', 'klem', 'study', 'ocampo_peñuela', 'en', 'knr']
92	6	0.825	result, study, sample, different, increase, model, value, effect, use, data	"*Rhizoctonia solani AG-1 IA infects both rice and signalgrass in the Colombian Llanos [Rhizoctonia solani AG-1 IA infecta arroz e braquiária nos Llanos Colombianos]*"	['rice', 'urochloa', 'isolate', 'solani_ag_ia', 'rhizoctonia', 'disease', 'urochloa_spp', 'plant', 'host', 'oryzae_sativae']
93	8	0.691	species, colombia, en, study, del, specimens, bird, record, include, base	"*Effect of wetland management: Are lentic wetlands refuges of plant-species diversity in the Andean-Orinoco Piedmont of Colombia?*"	['diversity', 'wetlands', 'species', 'wetland', 'plant', 'vegetation', 'landscape', 'lentic_wetlands', 'murillo_pacheco_peerj_peerj', 'plant_diversity']
94	6	0.949	result, study, sample, different, increase, model, value, effect, use, data	"*Instrumentation and control of a DC motor through the Ubidots platform*"	['control', 'devices', 'ubidots', 'allow', 'time', 'data', 'process', 'design', 'speed', 'controller']
95	6	0.982	result, study, sample, different, increase, model, value, effect, use, data	"*Short-term hyperprolactinemia decreases allergic inflammatory response of the lungs*"	['prolactin', 'group', 'cells', 'asthma', 'domperidone', 'il', 'effect', 'increase', 'rat', 'prl']
96	8	0.579	species, colombia, en, study, del, specimens, bird, record, include, base	"*Molecular phylogenetics of the white-lipped peccary (Tayassu pecari) did not confirm morphological subspecies in northwestern South America*"	['pecari', 'subspecies', 'pecari_albirostris', 'individuals', 'sequence', 'species', 'genetics_molecular_research_funpec', 'rp_funpecrp_com_br', 'populations', 'genetic_heterogeneity']
97	8	0.876	species, colombia, en, study, del, specimens, bird, record, include, base	"*A new species of tapaculo (Rhinocryptidae: Scytalopus) from the Serranía de Perijá of Colombia and Venezuela*"	['perijanus', 'note', 'species', 'latebricola', 'scytalopus', 'meridanus', 'serran_perija', 'colombia', 'icn', 'range']
98	0	0.82	case, value, orbit, order, correspond, present, point, energy, time, potential	"*Classical and quantum chaotic angular-momentum pumps*"	['spin', 'currents', 'angular_momentum', 'drive', 'scatter', 'chaotic', 'field', 'transport', 'current', 'incoming']
99	8	0.535	species, colombia, en, study, del, specimens, bird, record, include, base	"*Cultural tourism in Villavicencio Colombia*"	['villavicencio', 'llanero', 'tourism', 'dance', 'cultural', 'colombia', 'people', 'del', 'tourists', 'music']

Continúa en la siguiente página

Tabla B-2 – continuación de la página anterior

ID Doc.	Tema dom.	Prob. Tema	Términos Top-10 del tema	Título del documento	Términos Top-10 del Doc. según el Tema Dominante
100	6	0.968	result, study, sample, different, increase, model, value, effect, use, data	"*Spectrum occupancy measurement: An autocorrelation based scanning technique using USRP*"	['occupancy', 'result', 'channel', 'signal', 'mhz', 'band', 'gsm', 'value', 'energy_detection', 'compare']
101	8	0.858	species, colombia, en, study, del, specimens, bird, record, include, base	"*A distinctive new subspecies of scytalopus griseicollis (Aves, passeriformes, rhinocryptidae) from the northern eastern cordillera of Colombia and Venezuela*"	['griseicollis', 'morenoi', 'fte', 'icn', 'colombia_venezuela', 'specimens', 'species', 'colombia', 'subspecies', 'donegan_avendaño']
102	8	0.886	species, colombia, en, study, del, specimens, bird, record, include, base	"*'Bogotá' type specimens of the hummingbird genus Adelomyia, with diagnosis of an overlooked subspecies from the East Andes of Colombia*"	['specimens', 'melanogenys', 'type', 'santander_boyacá_population', 'adelomyia', 'sabinae', 'label', 'mnhn', 'note', 'bird']
103	6	0.982	result, study, sample, different, increase, model, value, effect, use, data	"*A method for medulloblastoma tumor differentiation based on convolutional neural networks and transfer learning*"	['feature', 'image', 'model', 'train', 'different', 'transfer_learn', 'ibca_cnn', 'obtain', 'approach', 'result']
104	8	0.906	species, colombia, en, study, del, specimens, bird, record, include, base	"*The birds of the Serranía de Perijá: The northernmost avifauna of the Andes [Las aves de la Serranía de Perijá: La avifauna más septentrional de los Andes]*"	['perijá', 'range', 'species', 'specimens', 'subspecies', 'record', 'bird', 'venezuela', 'serranía_perijá', 'cinco']
105	6	0.882	result, study, sample, different, increase, model, value, effect, use, data	"*The association between adult drug abuse and childhood maltreatment in students attending seven universities in five countries in Latin America and one country in the Caribbean [A associacao entre o abuso de drogas e o maltrato infantil em estudantes de sete universidades de cinco paises da América Latina e um do Caribe] [La asociacion entre el abuso de drogas y el maltrato infantil en estudiantes de siete uniersidades de cinco paises de Latin America y uno pais del Caribe]*"	['link', 'drug_abuse', 'study', 'childhood_maltreatment', 'report', 'maltreatment', 'sample', 'drug', 'maltrato', 'students']
106	6	0.854	result, study, sample, different, increase, model, value, effect, use, data	"*Climate-smart crop-livestock systems for smallholders in the tropics: Integration of new forage hybrids to intensify agriculture and to mitigate climate change through regulation of nitrification in soil*"	['bni', 'hybrids', 'humidicola', 'nitrification', 'subbarao', 'ciat', 'rao', 'soil', 'new', 'output']
107	0	0.87	case, value, orbit, order, correspond, present, point, energy, time, potential	"*Optimal conditions for the numerical calculation of the largest Lyapunov exponent for systems of ordinary differential equations*"	['lle', 'value', 'order', 'particle_method', 'method', 'orbit', 'du_ng', 'lorenz', 'result', 'renormalization_time']
108	6	0.899	result, study, sample, different, increase, model, value, effect, use, data	"*Orientation (Yaw) Fuzzy controller applied to a car-like mobile robot prototype*"	['vehicle', 'value', 'output', 'error', 'helvis', 'input', 'controller', 'speed', 'steer', 'set']
109	6	0.691	result, study, sample, different, increase, model, value, effect, use, data	"*Molecular characterization of 93 genotypes of cocoa (Theobroma cacao L.) with random amplified microsatellites RAMs [Caracterización molecular con microsatélites amplificados al azar (RAMs) de 93 genotipos de cacao (Theobroma cacao L.)]*"	['genotypes', 'genetic_diversity', 'cocoa', 'tumaco', 'group', 'san_luis_robles', 'ram', 'high', 'theobroma_cacao', 'study']
110	8	0.85	species, colombia, en, study, del, specimens, bird, record, include, base	"*A new Tapaculo related to Scytalopus rodriguezi from Serranía de los Yariguíes, Colombia*"	['rodriguezi', 'species', 'record', 'note', 'yariguiorum', 'song', 'donegan_avendaño', 'tapaculo', 'yariguíes', 'colombia']
111	6	0.697	result, study, sample, different, increase, model, value, effect, use, data	"*Quantitative measurement of epithelial injury in ornamental silver dollar fish (Metynnis orinocensis) captured in the wild, imported wild-caught, and aquacultured*"	['handle', 'injury', 'skin', 'ornamental_fish', 'wild', 'injuries', 'fluorescein', 'skin_injuries', 'aquaculture', 'net']

Continúa en la siguiente página

Tabla B-2 – continuación de la página anterior

ID Doc.	Tema dom.	Prob. Tema	Términos Top-10 del tema	Título del documento	Términos Top-10 del Doc. según el Tema Dominante
112	6	0.784	result, study, sample, different, increase, model, value, effect, use, data	"*Characterization and functional ecology of the extrafloral nectar of Cedrela odorata in contrasting growth environments in central Veracruz, Mexico*"	['plant', 'efn', 'odorata', 'nectar', 'page', 'ants', 'sugar', 'ant', 'water', 'lm']
113	6	0.54	result, study, sample, different, increase, model, value, effect, use, data	"*Evidence of small modulation of ethinylestradiol induced effects by concurrent exposure to trenbolone in male eelpout Zoarces viviparus*"	['ee', 'tb', 'testis', 'ng', 'severe', 'control', 'er', 'low', 'high', 'group']
114	8	0.596	species, colombia, en, study, del, specimens, bird, record, include, base	"*Seasonal variation and natural infection of lutzomyia antunesi (Diptera: Psychodidae: Phlebotominae), an endemic species in the orinoquia region of Colombia*"	['lutzomyia', 'lu_antunesi', 'abundance', 'species', 'colombia', 'areas', 'leishmania', 'diptera_psychodidae', 'natural_infection', 'acl']
115	6	0.922	result, study, sample, different, increase, model, value, effect, use, data	"*Exploring child maltreatment and its relationship to alcohol and cannabis use in selected Latin American and Caribbean countries*"	['child_maltreatment', 'maltreatment', 'measure', 'report', 'psychological_distress', 'abuse', 'result', 'research', 'substance_use', 'religiosity']
116	8	0.741	species, colombia, en, study, del, specimens, bird, record, include, base	"*Molecular phylogeny and redefined generic limits of Calathea (Marantaceae)*"	['species', 'calathea', 'clade', 'schumann', 'genus', 'include', 'section', 'kennedy', 'sua_rez_col_jq', 'sequence']
117	6	0.763	result, study, sample, different, increase, model, value, effect, use, data	"*Use of sodium chloride and zeolite during shipment of Ancistrus triradiatus under high temperature*"	['shipment', 'zeolite', 'transport', 'mortality', 'salt', 'group', 'triradiatus', 'stress', 'total_ammonia', 'aquaculture']
118	6	0.8	result, study, sample, different, increase, model, value, effect, use, data	"*Bezafibrate, a lipid-lowering pharmaceutical, as a potential endocrine disruptor in male zebrafish (Danio rerio)*"	['cholesterol', 'beza_brate', 'expression', 'level', 'genes', 'kt', 'aquarium', 'bzf', 'nest_anova', 'sample']
119	6	0.969	result, study, sample, different, increase, model, value, effect, use, data	"*Dual effects of hyperprolactinemia on carrageenan-induced inflammatory paw edema in rats*"	['prl', 'prolactin', 'rat', 'animals', 'effect', 'day', 'hyperprolactinemia', 'domperidone', 'volume', 'injection']
120	6	0.866	result, study, sample, different, increase, model, value, effect, use, data	"*Cryopreservation effects on the sperm quality of cachama blanca Piaractus brachypomus (Cuvier 1818)*"	['straw', 'sperm', 'fertility', 'thaw', 'ml', 'cryopreservation', 'meet', 'dmso', 'etg', 'evaluate']
121	6	0.486	result, study, sample, different, increase, model, value, effect, use, data	"*Fish farming of native species in Colombia: Current situation and perspectives*"	['cruz_casallas', 'production', 'aquaculture', 'farm', 'food', 'lo_pez', 'nal', 'tion', 'growth', 'tilapia']
122	1	0.556	fish, sh, observe, exposure, water, effect, concentration, endosulfan, gill, liver	"*Endosulfan affects health variables in adult zebrafish (Danio rerio) and induces alterations in larvae development*"	['endosulfan', 'sh', 'exposure', 'effect', 'observe', 'liver', 'mg', 'expose_endosulfan', 'day', 'na_atpase_activity']
123	0	0.908	case, value, orbit, order, correspond, present, point, energy, time, potential	"*The influence of the Lande g-factor in the classical general relativistic description of atomic and subatomic systems*"	['corrections', 'potentials', 'order', 'proton', 'electron', 'classical', 'spin', 'phys_rev', 'iq', 'value']
124	6	0.933	result, study, sample, different, increase, model, value, effect, use, data	"*Short distance movement of genomic negative strands in a host and nonhost for Sugarcane mosaic virus (SCMV)*"	['plant', 'virus', 'viral', 'scmv', 'maize', 'movement', 'resistance', 'stem', 'leave', 'rna']
125	6	0.764	result, study, sample, different, increase, model, value, effect, use, data	"*Nutritional requirements of freshwater ornamental fish: A review*"	['diet', 'ornamental_fish', 'food', 'growth', 'protein', 'source', 'requirements', 'feed', 'vitamin', 'dietary_protein']
126	6	0.91	result, study, sample, different, increase, model, value, effect, use, data	"*Child Maltreatment and Its Relationship to Drug Use in Latin America and the Caribbean: An Overview and Multinational Research Partnership*"	['child_maltreatment', 'drug', 'report', 'drug_use', 'use', 'children', 'abuse', 'sexual_abuse', 'case', 'child']
127	0	0.876	case, value, orbit, order, correspond, present, point, energy, time, potential	"*Innermost stable circular orbits around magnetized rotating massive stars*"	['magnetic_field', 'mass', 'magnetic_dipole', 'isco', 'case', 'neutron_star', 'radius_isco', 'approximate', 'magnetars', 'formula']
128	0	0.831	case, value, orbit, order, correspond, present, point, energy, time, potential	"*Solving the time-dependent schrödinger equation with absorbing boundary conditions and source terms in mathematica 6.0*"	['numerical', 'source_term', 'eq', 'case', 'matrix', 'grid', 'method', 'finite_square', 'representation', 'quantum_mechanics']

Continúa en la siguiente página

Tabla B-2 – continuación de la página anterior

ID Doc.	Tema dom.	Prob. Tema	Términos Top-10 del tema	Título del documento	Términos Top-10 del Doc. según el Tema Dominante
129	6	0.976	result, study, sample, different, increase, model, value, effect, use, data	"*Acute and chronic stress and the inflammatory response in hyperprolactinemic rats*"	['stress', 'prl', 'group', 'domperidone', 'acute_stress', 'effect', 'cold_stress', 'induce', 'animals', 'vehicle']
130	6	0.524	result, study, sample, different, increase, model, value, effect, use, data	"*Behavioural and gill histopathological effects of acute exposure to sodium chloride in moneda (Metynnis orinocensis)*"	['salt', 'metynnis_orinocensis', 'nacl', 'mortality', 'salinity', 'lc', 'respectively', 'code', 'juveniles', 'sodium_chloride']
131	6	0.866	result, study, sample, different, increase, model, value, effect, use, data	"*Seasonal variation of sperm quality and the relationship between spermatocrit and sperm concentration in yamú Brycon amazonicus*"	['reproductive_season', 'seminal_plasma', 'semen', 'yamu', 'periods', 'sperm_concentration', 'value', 'period', 'determine', 'glucose']
132	6	0.908	result, study, sample, different, increase, model, value, effect, use, data	"*Cryopreservation of yamú (Brycon amazonicus) sperm for large scale fertilization*"	['straw', 'sperm', 'semen', 'egg', 'thaw', 'cryopreservation', 'fertility', 'fertilization', 'ml', 'spermatozoa']
133	6	0.748	result, study, sample, different, increase, model, value, effect, use, data	"*Milt quality and spermatozoa morphology of captive Brycon siebenthalae (Eigenmann) broodstock*"	['sperm', 'semen', 'spermatozoa', 'egg', 'aquaculture', 'sperm_motility', 'spermatocrit', 'ml', 'sample', 'motility']
134	6	0.694	result, study, sample, different, increase, model, value, effect, use, data	"*The role of livestock in poverty alleviation: An analysis of Colombia*"	['antioquia', 'smallholders', 'regions', 'cundiboyacense_altiplanicie', 'farmers', 'interview', 'producers', 'poverty', 'coffee', 'piedmont_caribbean']
135	6	0.963	result, study, sample, different, increase, model, value, effect, use, data	"*Reproductive experience modulates dopamine-related behavioral responses*"	['amph', 'rat', 'da', 'induce', 'effect', 'min', 'behavioral', 'dopaminergic', 'behavior', 'females']
136	6	0.985	result, study, sample, different, increase, model, value, effect, use, data	"*A quantitative analysis of the role of experience in the regulation of sexual behavior in male rats*"	['test', 'sexual_behavior', 'ejaculation', 'behavior', 'influence', 'train', 'effect', 'male_rat', 'experience', 'parameters']

C. Anexo: Reporte de la producción académica de la Universidad de los Llanos

La producción académica-científica de las Instituciones de Educación Superior (IES) como tesis, informes finales de trabajo de grado, y principalmente, artículos científicos es el resultado de investigaciones y fuente principal de generación de nuevo conocimiento de alto nivel, siendo las bases en el desarrollo tecnológico y académico para la innovación y mejora en el ámbito económico y productivo. Las IES, como la Universidad de los Llanos, requieren identificar sus campos de acción e investigación y así conocer su impacto en la sociedad, este análisis se puede realizar de forma manual, requiriendo inversiones y recursos adicionales constantemente, o automática, que usa eficientemente menos recursos. La forma automática puede ser implementada por métodos o herramientas externas o internas. Este trabajo, motivado por el uso de recursos internos en pro de dinamizar la economía local y la adopción tecnológica, realizó un estudio de la producción académica-científica de la Universidad de los Llanos (Unillanos), para la identificación temática y áreas de investigación en artículos científicos publicados e indexados en SCOPUS por autores con filiación a la universidad. Los artículos que fueron analizados corresponden a los publicados entre 1999 y el 28 de octubre de 2019 (20 años) de los cuales se obtuvieron 293 documentos escritos en varios idiomas (ver figura **C-1**), siendo los dominantes Inglés y Español con 137 y 113, respectivamente. Este trabajo analizó únicamente los documentos escritos en Inglés debido al alcance del mismo. La distribución de estos documentos por áreas se muestra en la figura **C-2**. Un cambio que cabe resaltar es el área de "Veterinaria" ("*Veterinary*") la cual reduce significativamente la cantidad de documentos pertenecientes a dicha área. Esto se debe a que los documentos publicados clasificados en esta área son escritos principalmente en Español. Respecto a los documentos en Inglés, se observa que las áreas de sector Agropecuario y ciencias biológicas en conjunto con el sector ingenieril y ciencias de la computación son las de mayor producción académico-científico en Inglés, idioma de mayor preferencia por aumentar el impacto y alcance investigativo. Esto permite dar una primera idea de los campos de investigación de la Unillanos.

Este trabajo identificó automáticamente 10 temas a partir de la producción científica, algunos de estos temas más relevantes que otros. Cada tema conformado por una distribución

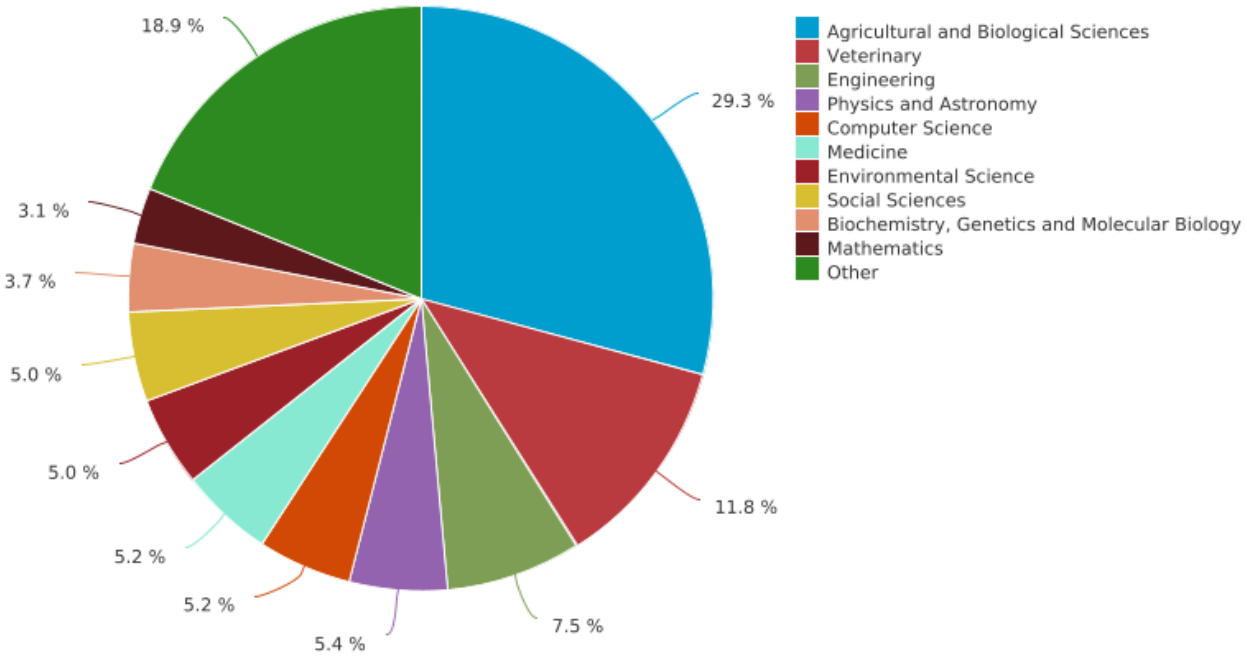


Figura C-1.: Distribución de 293 artículos científicos por áreas según SCOPUS. Adaptado de SCOPUS.

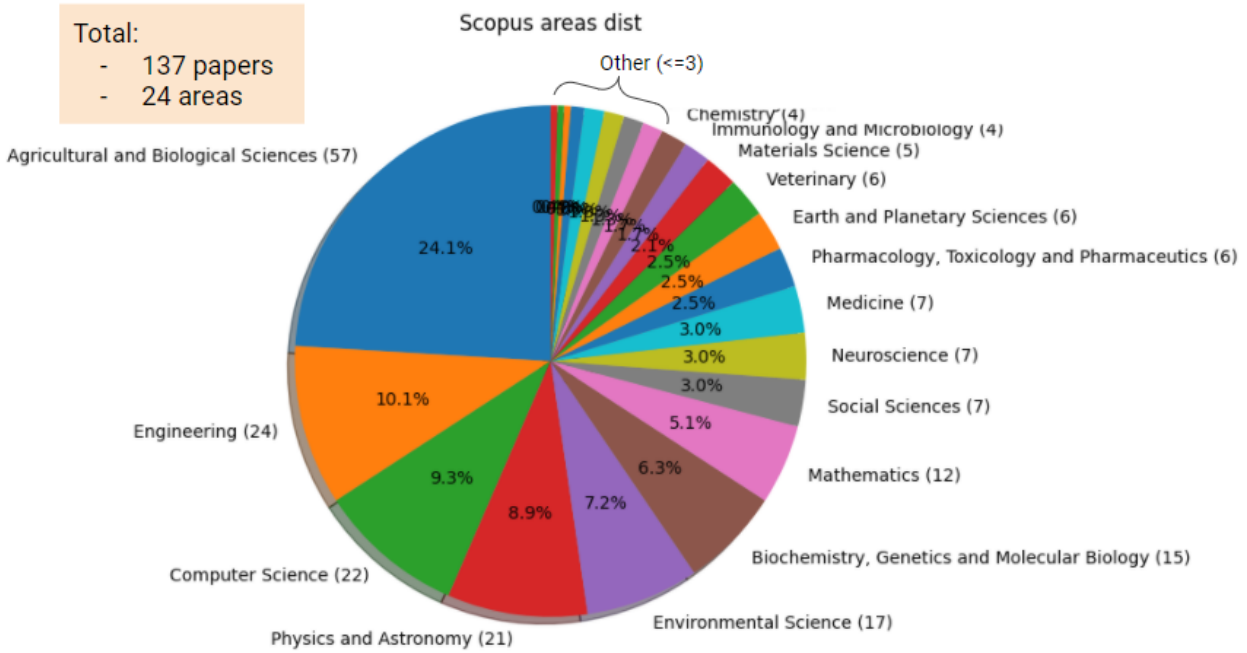


Figura C-2.: Distribución de 137 artículos científicos en inglés por áreas según SCOPUS. Elaboración propia.

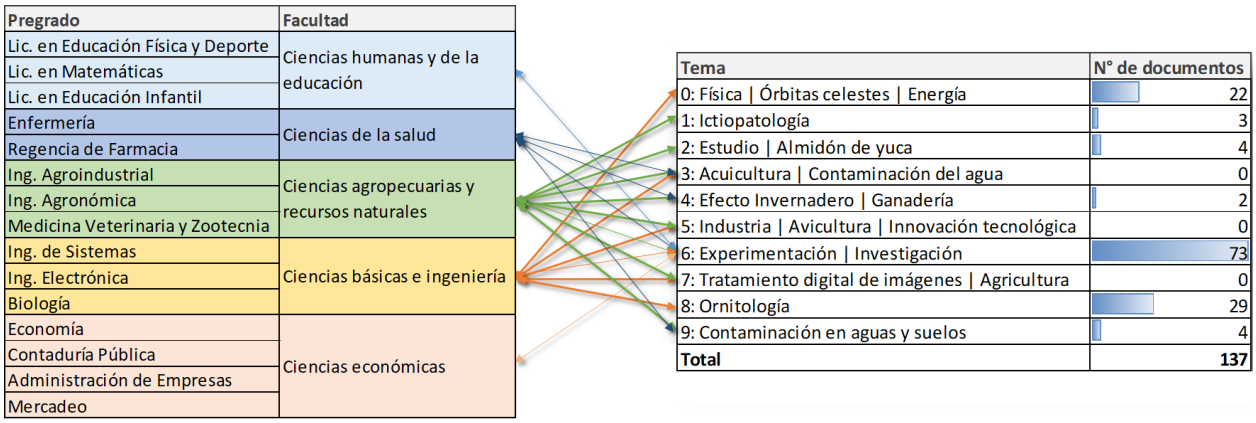
probabilística de términos (palabras) provenientes del conjunto de datos. A partir del análisis de estos términos se describió y denominó cada uno de los temas.

Desde su fundación, Unillanos ha tenido diversos programas académicos siendo que algunos fueron reformados, dejaron de ofertarse o se crearon. Unillanos, a la fecha de realización de este trabajo y escritura de este informe, tiene 19 programas de pregrado, 16 especializaciones, 10 maestrías y 1 doctorado (programas de educación superior e incluyendo todas las sedes) agrupados en 5 facultades. Con fines prácticos y para facilitar la interpretación de los resultados de este trabajo, se incluyó implícitamente los programas de posgrado a los programas de pregrado, considerando que dichos posgrados corresponden a la misma linea de estudios de algunos programas de pregrado. Recientemente, fueron ofertados dos nuevos programas por la Universidad (Ingeniería de Procesos e Ingeniería Ambiental) por lo que son descartados en este análisis.

Otro factor importante que cabe resaltar, es la relación atemporal que se realiza en este trabajo entre los temas y los programas establecidos, es decir, no se establecieron relaciones con programas dejaron de existir durante el periodo establecido. Esto se debe al foco de investigación y alcance de este estudio. Las relaciones se hicieron tomando en cuenta los programas establecidos al considerarse un cambio poco significativo históricamente de los programas y campos de investigación de la Universidad.

Se identificaron 10 temas automáticamente a partir de los 137 artículos científicos escritos en inglés los cuales fueron asociados a las facultades de Unillanos. Estas asociaciones se limitaron a las facultades dado que el alcance de este trabajo no permitió asociar, en un nivel más específico, con cada uno de los programas de pregrado. Estas asociaciones se muestran en la figura **C-3** por medio de flechas donde cada facultad tiene asociado un color distintivo usado igualmente en las flechas. Adicionalmente, se muestra el número de documentos que consideran dicho tema como principal o dominante. Se estableció dos niveles de asociación: *i)* una asociación "fuerte" que se entiende como una relación clara y directa entre la facultad (o algunos de sus programas) y el tema; y *ii)* una asociación "suave" que señala una relación poco clara o indirecta. Estas asociaciones fueron establecidas según análisis, deducción y experticia de los autores de este trabajo.

A destacar, las asociaciones establecidas del tema 6 "Experimentación |Investigación" las cuales se conectan con todas las facultades, esto se debe a que dicho tema referencia todos al proceso de investigación en sí sin definir claramente el objetivo u problema abordado, adicionalmente, el contexto de los documentos procesados, siendo artículos científicos, corresponden a producción científica a través de procesos de investigación y experimentación. Por lo dicho anteriormente, se establece que el tema 6 se relaciona con el concepto de investigación y experimentación, procesos transversales en las áreas y facultades. Esto se corrobora



Pregrado	Facultad
Lic. en Educación Física y Deporte	Ciencias humanas y de la educación
Lic. en Matemáticas	
Lic. en Educación Infantil	
Enfermería	Ciencias de la salud
Regencia de Farmacia	
Ing. Agroindustrial	Ciencias agropecuarias y recursos naturales
Ing. Agronómica	
Medicina Veterinaria y Zootecnia	
Ing. de Sistemas	Ciencias básicas e ingeniería
Ing. Electrónica	
Biología	
Economía	Ciencias económicas
Contaduría Pública	
Administración de Empresas	
Mercadeo	

Tema	N° de documentos
0: Física | Órbitas celestes | Energía	22
1: Ictiopatología	3
2: Estudio | Almidón de yuca	4
3: Acuicultura | Contaminación del agua	0
4: Efecto Invernadero | Ganadería	2
5: Industria | Avicultura | Innovación tecnológica	0
6: Experimentación | Investigación	73
7: Tratamiento digital de imágenes | Agricultura	0
8: Ornitología	29
9: Contaminación en aguas y suelos	4
Total	**137**

Figura C-3.: Asociación de facultades de Unillanos y 10 temas identificados según la producción académica-científica. Elaboración propia.

con el número de documentos del tema 6.

Los temas de mayor impacto, omitiendo el tema 6 por lo mencionado anteriormente, son el 8 "Ornitología" y 0 "Física | Órbitas celestes | Energía". Por el contrario, los temas de menor producción son el 3 "Acuicultura | Contaminación del agua", 5 "Industria | Avicultura | Innovación tecnológica" y 7 "Tratamiento digital de imágenes | Agricultura".

Otro característica importante es la gran cantidad de temas asociados a las facultades "Ciencias agropecuarias y recursos naturales" (FCARN) y "Ciencias básicas e ingeniería" (FCBI), siendo las facultades con más temas asociados: Seguidamente, la facultad de "Ciencias de la salud" (FCS) con algunas asociaciones "suaves". Por último, las facultades de "Ciencias humanas y de la educación" (FCHYE) y "Ciencias económicas" (FCE) con una sola asociación "suave". Esto permite concluir que: *i)* Las facultades con mayor producción académico-científico en Inglés y clasificables en temas diferentes son FCARN y FCBI, *ii)* FCS no tiene temas directos definidos, *iii)* Las facultades con menor producción académico-científico son FCHYE y FCE, esto posiblemente por el idioma en que están escritos los documentos (e.g. Español) y el alcance de los mismos.

Se recomienda a la universidad dinamizar la adopción tecnológica por medio de la implementación y el desarrollo de herramientas tecnológicas internas o locales. Igualmente, promover la producción académico-científico en las temas más rezagados y seguir fortaleciendo aquellos que ya tienen un alto impacto. Además, de incentivar la investigación para la generación de nuevo conocimiento por medio de la escritura de artículos científicos en Inglés, tanto en los programas de pregrado que presentaron mayor producción como aquellos que no, dado que para la Unillanos, siendo una universidad de alto impacto en la región, se considera un poco bajo el número de publicaciones encontradas.

D. Anexo: Número de evaluaciones cualitativas por cada artículo científico

En las figuras D-1 y D-2 se contabilizan el número de evaluaciones hechas por cada experto en cada artículo científico para las evaluaciones de valoración de la asociatividad (V2) y valoración por grado de representatividad (V3), respectivamente.

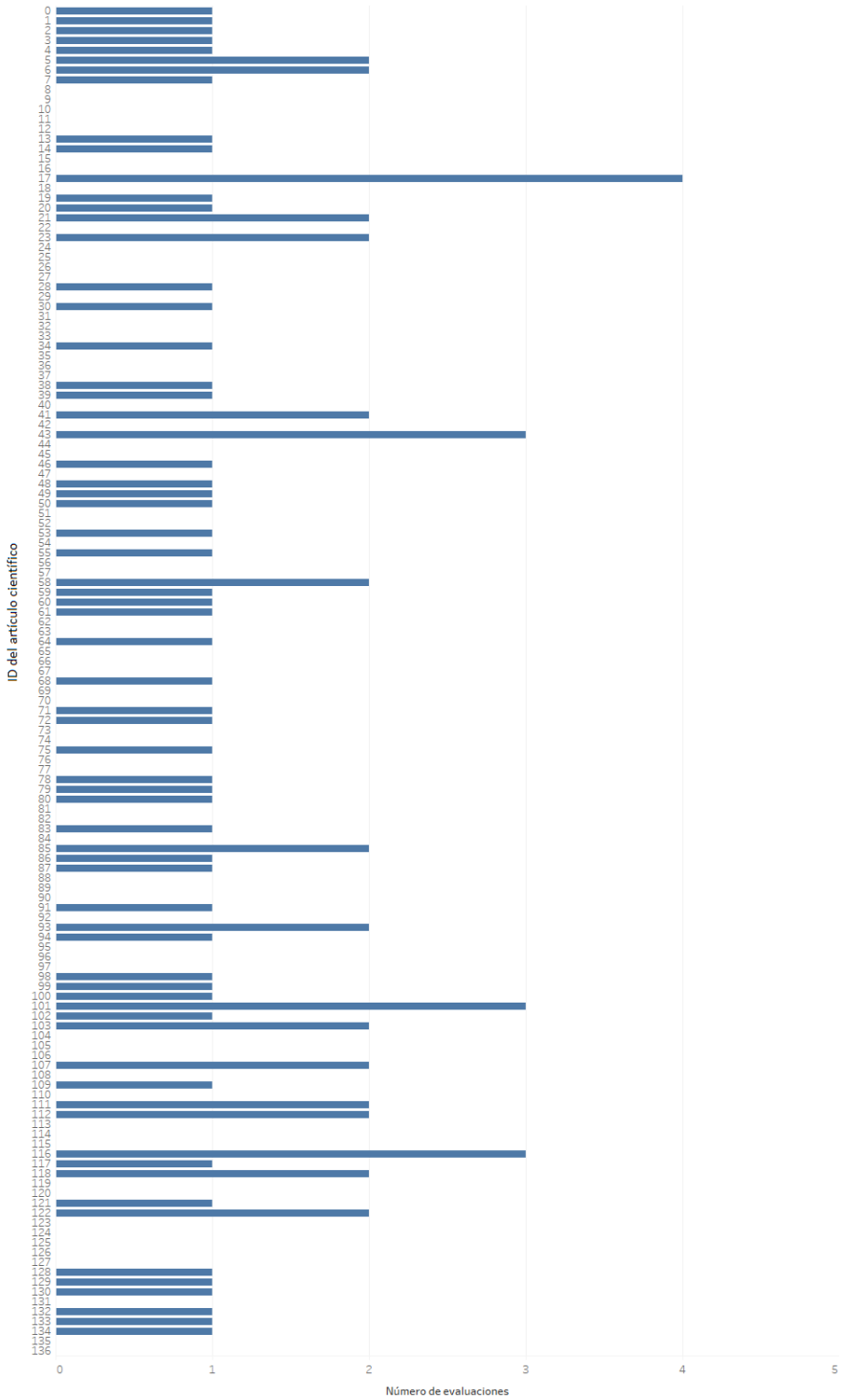


Figura D-1.: Cantidad de evaluaciones registradas por artículo por los expertos para la valoración de la asociatividad (V2). Elaboración propia.

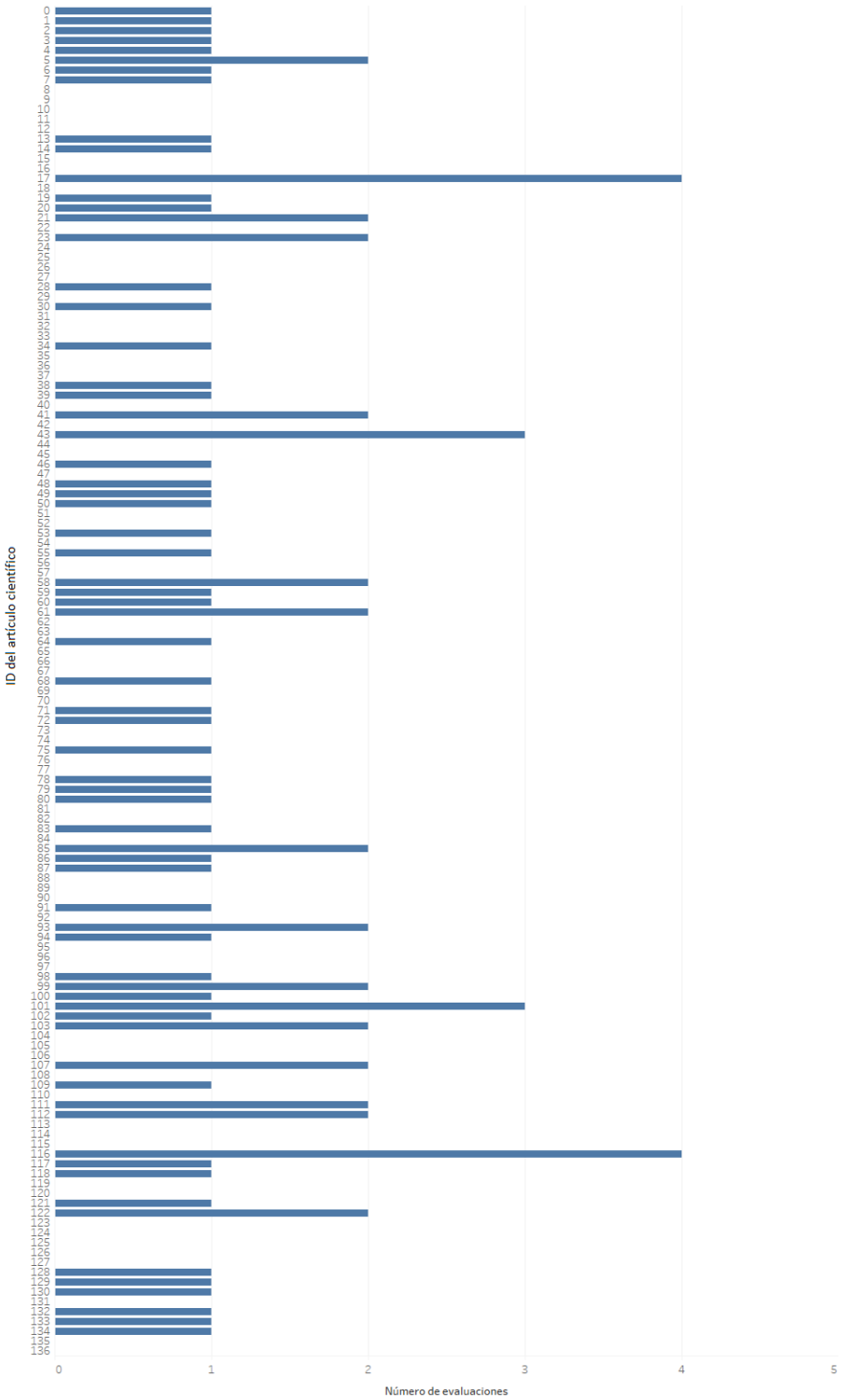


Figura D-2.: Cantidad de evaluaciones registradas por artículo por los expertos para la valoración por grado de representatividad (V3). Elaboración propia.

E. Anexo: Número de evaluaciones cualitativas por cada documento vs el nivel de coherencia

En las figuras **E-1** y **E-2** se contabilizan el número de evaluaciones de cada artículo científico por grado de coherencia determinado por cada experto para las evaluaciones de valoración de la asociatividad (V2) y valoración por grado de representatividad (V3), respectivamente.

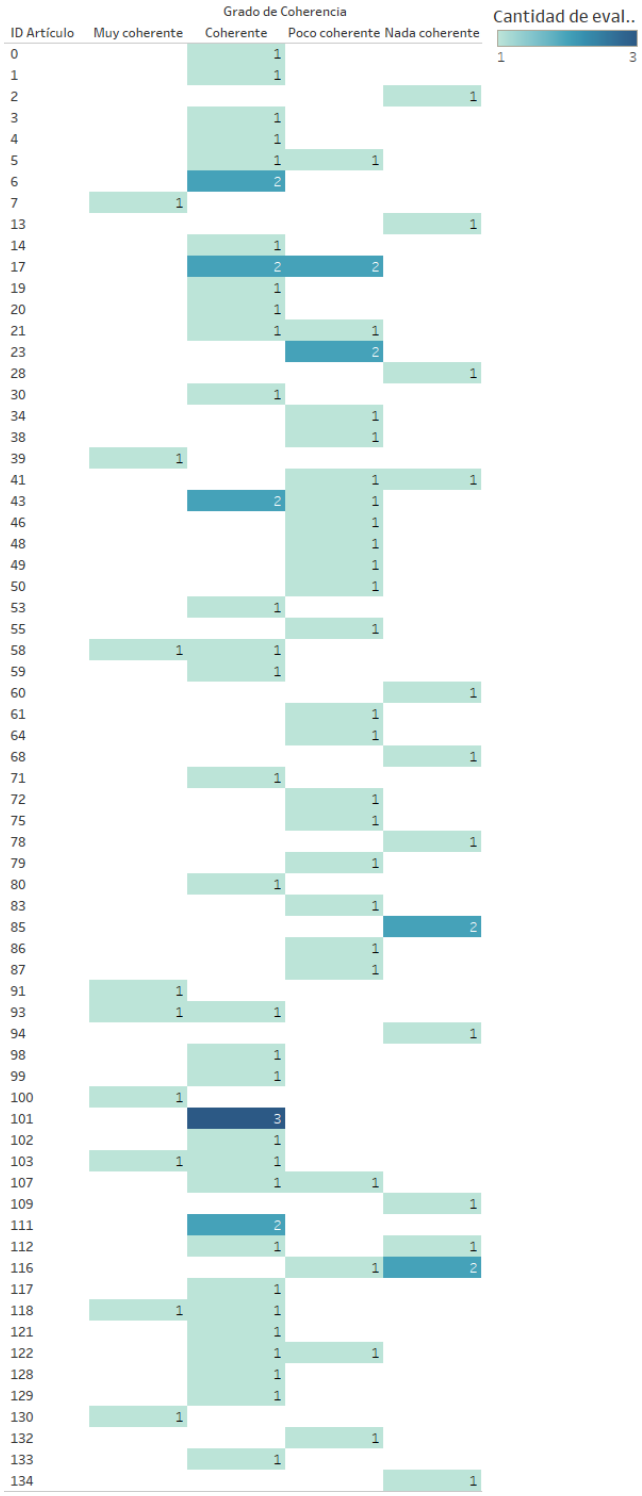


Figura E-1.: Cantidad de evaluaciones registradas por artículo según el nivel de coherencia asignada por los expertos para la V2. Elaboración propia.

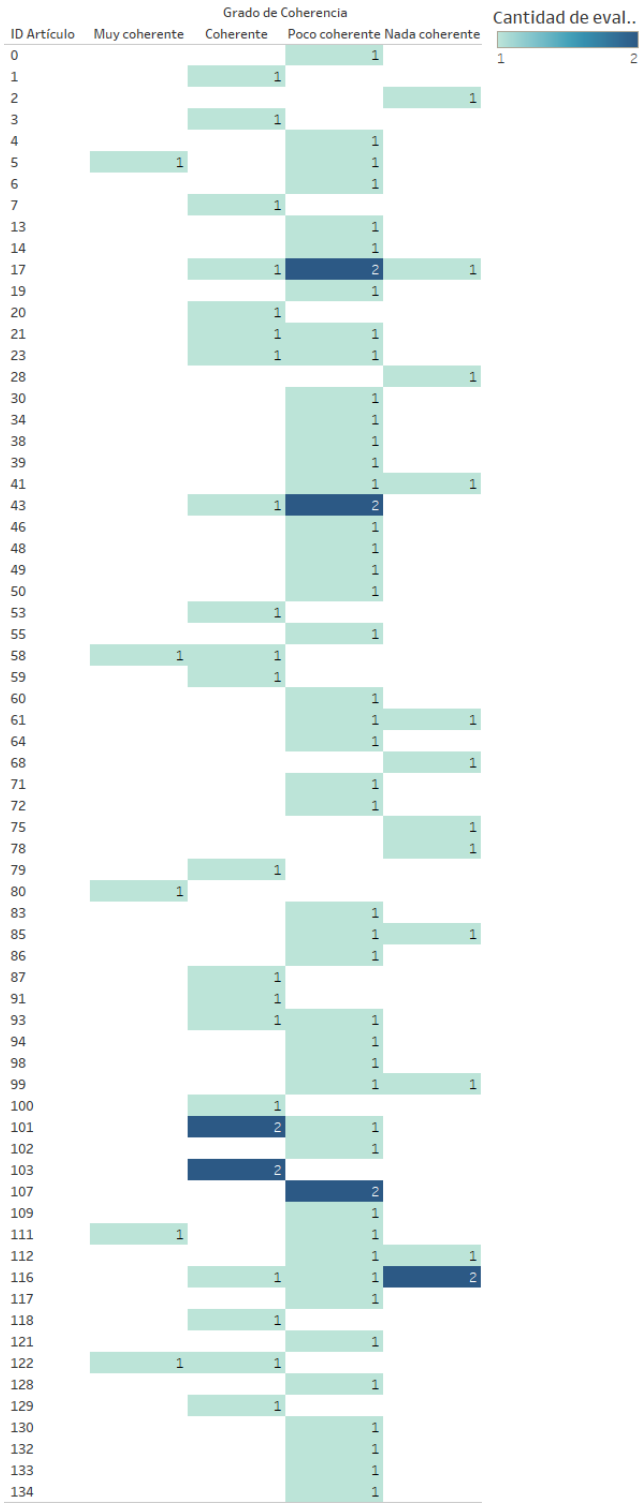


Figura E-2.: Cantidad de evaluaciones registradas por artículo según el nivel de coherencia asignada por los expertos para la V3. Elaboración propia.

F. Anexo: Descripción del proceso de diseño e implementación de la aplicación web whatTopic

Como herramienta/instrumento para la evaluación cualitativa se propuso diseñar, crear y desplegar una aplicación web que permitiera a los expertos acceder y diligenciar dicha evaluación. Esta herramienta fue denominada "whatTopic". El proceso de diseño y construcción de "whatTopic" se describe brevemente a continuación:

1. Versión 0.0.0: En esta etapa se diseñaron unos *mockups* que sirvieron como base para estructurar el instrumento, véase la figura **F-1**.

2. Versión 0.0.1: Corresponde al código base realizado en Django con la estructura inicial según los *mockups*, véase la figura **F-2**.

3. Versión 0.9.0: La estructura de las tres (3) evaluaciones es creada.

4. Versión 1.0.0: Se termina la estructura de las tres (3) evaluaciones y se redacta el texto introductorio del instrumento y la descripción y actividad en cada evaluación. En la figura **F-4** se puede ver la evaluación V3.

5. Versión 1.4.0: Es la última versión y la desplegada en internet para el desarrollo de la actividad de evaluación por parte de los expertos. Su interfaz gráfica no varía mucho de la v1.0.0. Sus cambios son principalmente de funcionalidad y redacción. La sección introductoria se muestra en la figura **F-5**, la sección de la evaluación V1 en **F-6**, la sección de la evaluación V2 en **F-7**, la sección de la evaluación V3 en **F-8**, la sección de agradecimientos en **F-9** y la sección *Footer* en **F-10**.

6. El modelo entidad/relación de la base de datos implementada en la versión 1.4.0 se puede observar en la figura **F-11**.

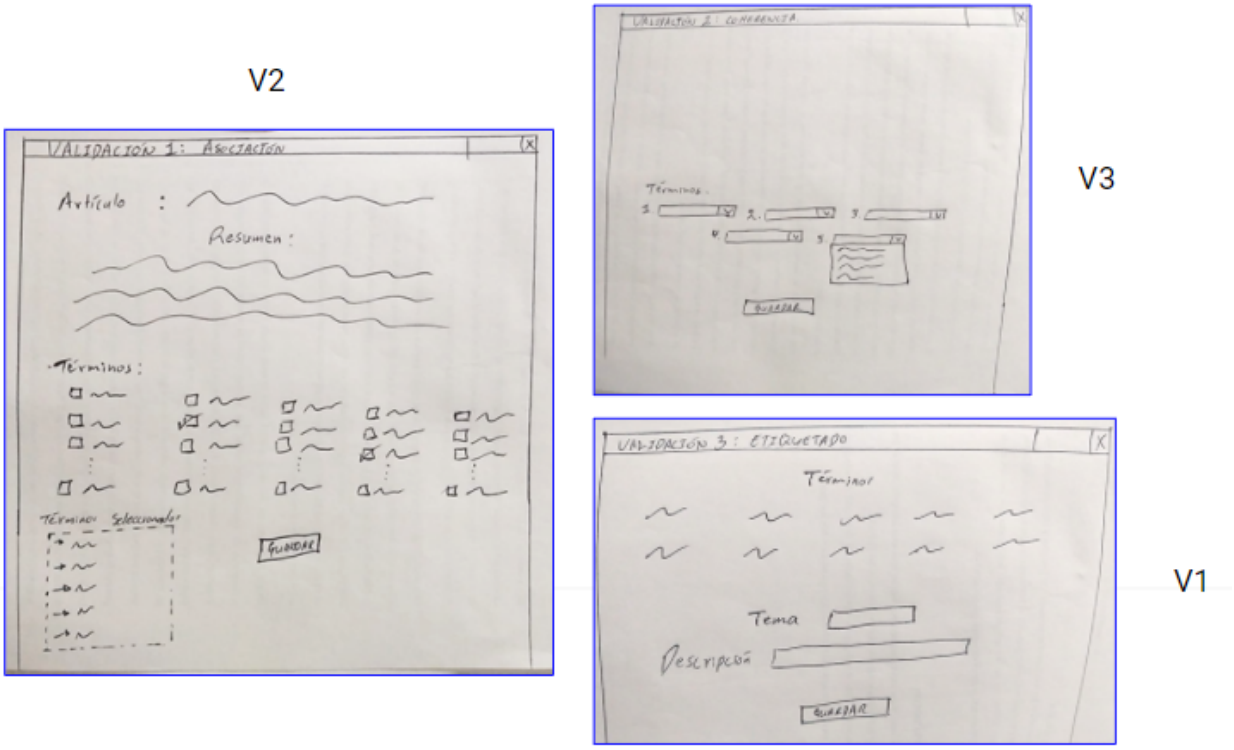


Figura F-1.: "whatTopic" v0.0.0. Elaboración propia.

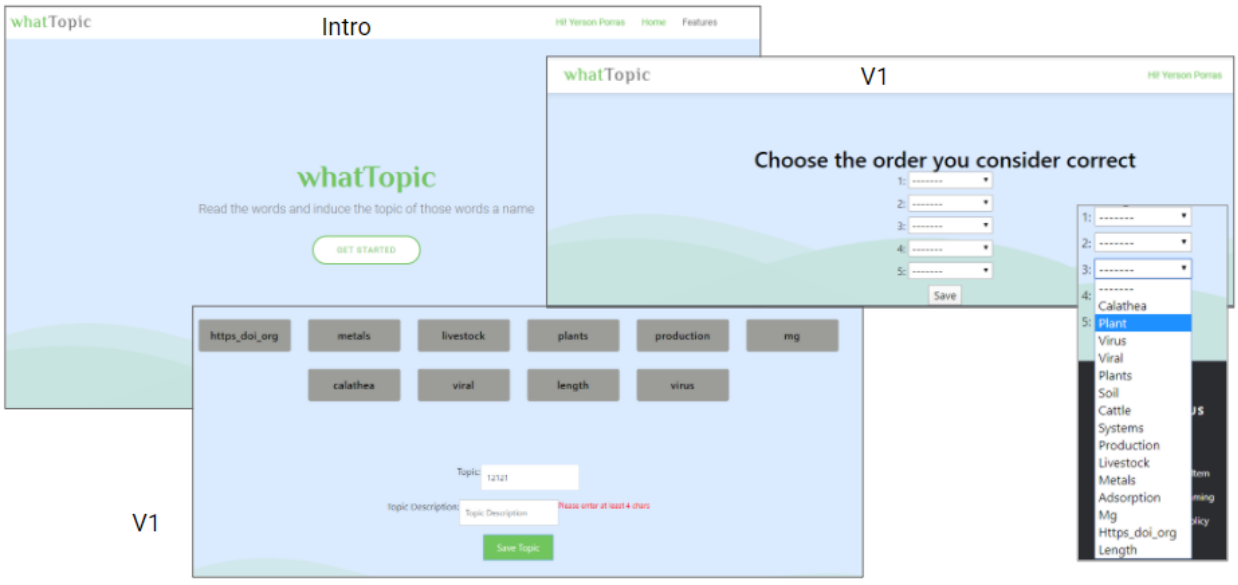


Figura F-2.: "whatTopic" v0.0.1. Elaboración propia.

Figura F-3.: "whatTopic" v0.9.0. Elaboración propia.

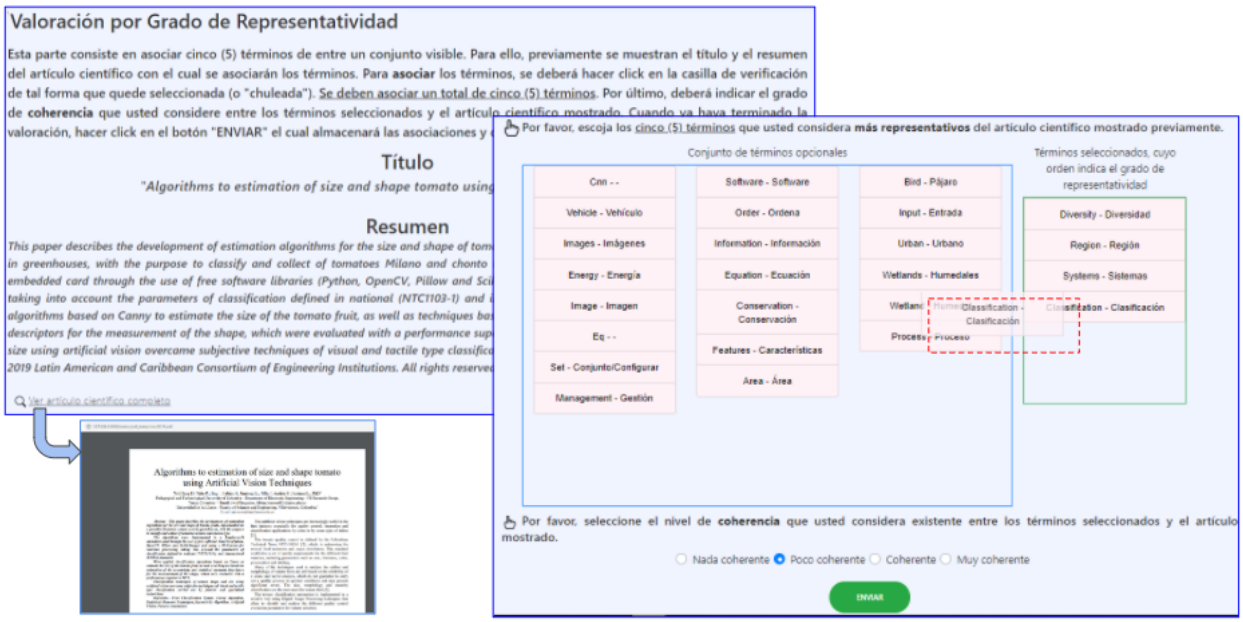


Figura F-4.: "whatTopic" v1.0.0. Fragmento V3. Elaboración propia.

Figura F-5.: "whatTopic" v1.4.0. Sección Introductoria. Elaboración propia.

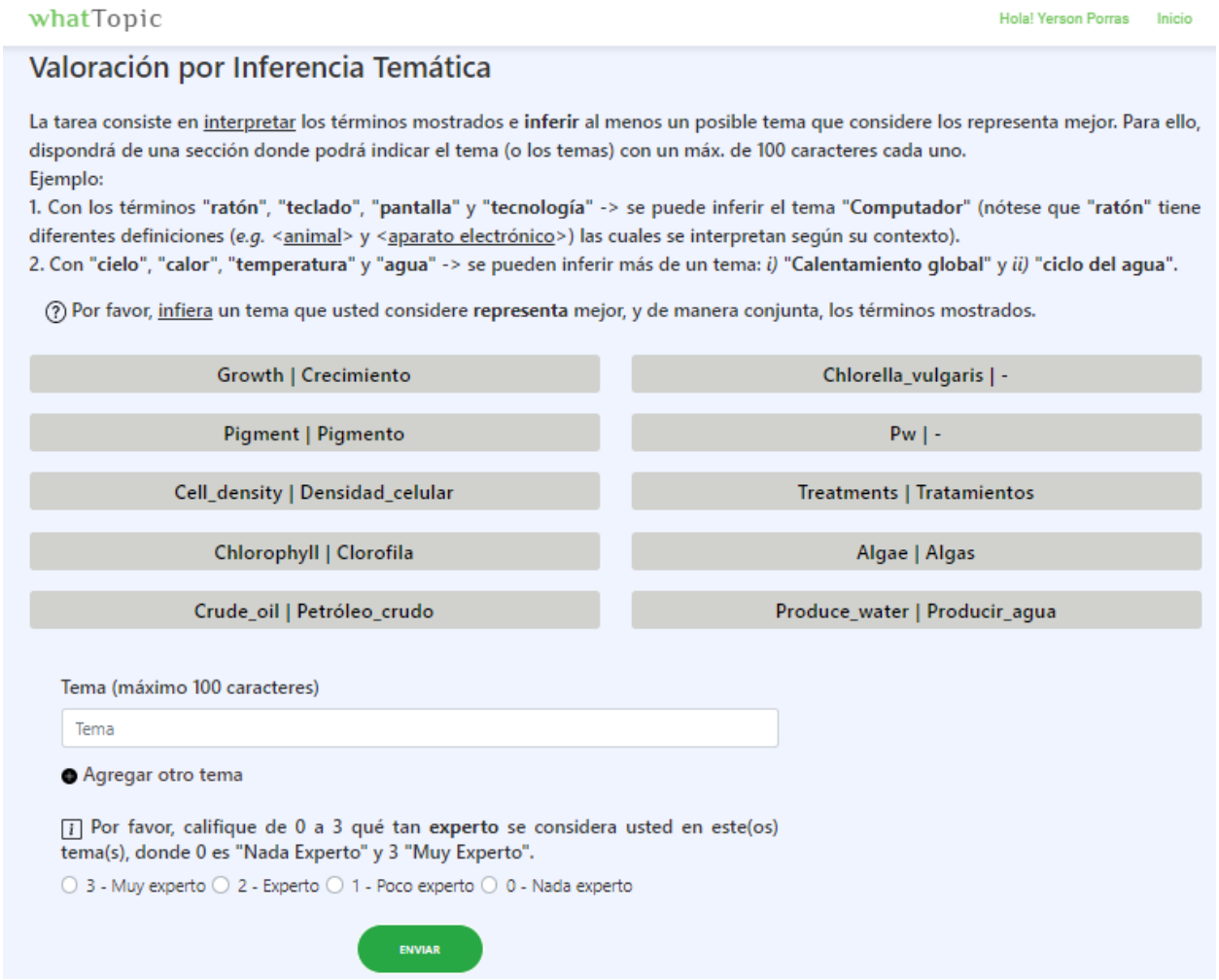


Figura F-6.: "whatTopic" v1.4.0. Sección Evaluación V1. Elaboración propia.

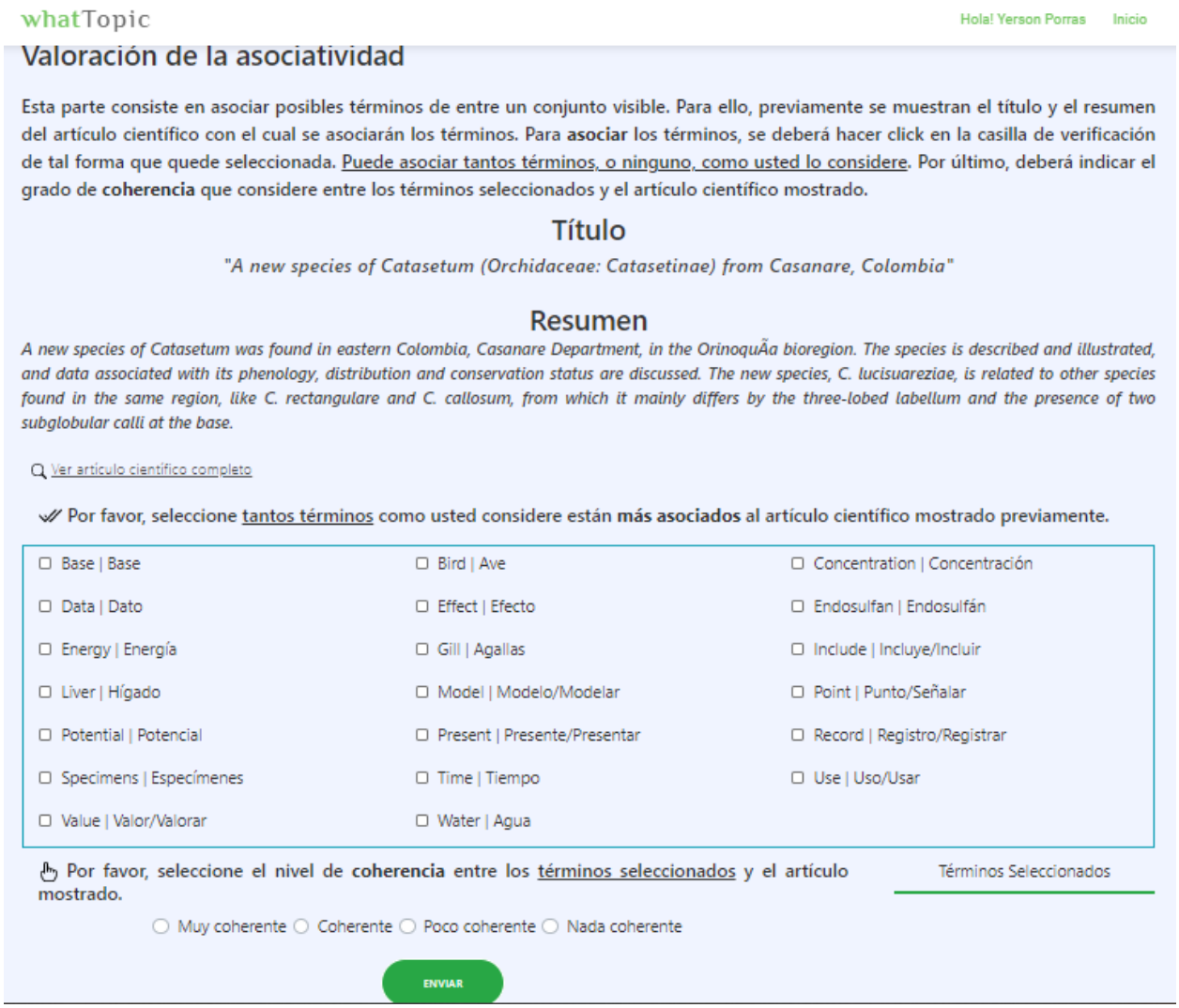


Figura F-7.: "whatTopic" v1.4.0. Sección Evaluación V2. Elaboración propia.

Figura F-8.: "whatTopic" v1.4.0. Sección Evaluación V3. Elaboración propia.

Figura F-9.: "whatTopic" v1.4.0. Sección Agradecimientos. Elaboración propia.

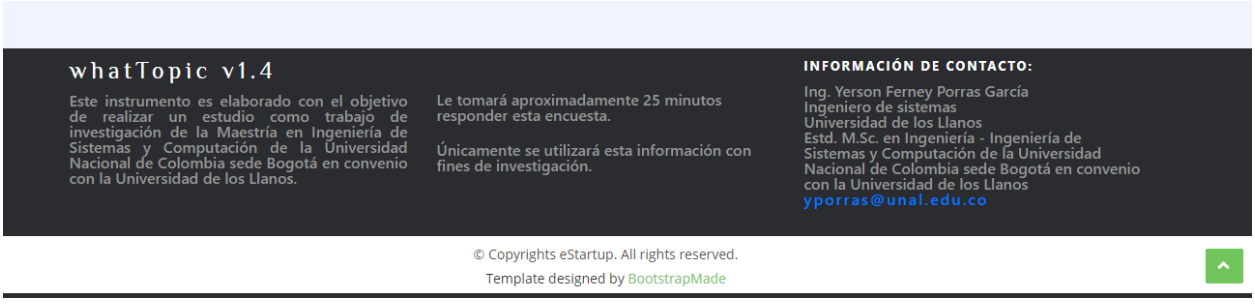


Figura F-10.: "whatTopic" v1.4.0. Sección Footer. Elaboración propia.

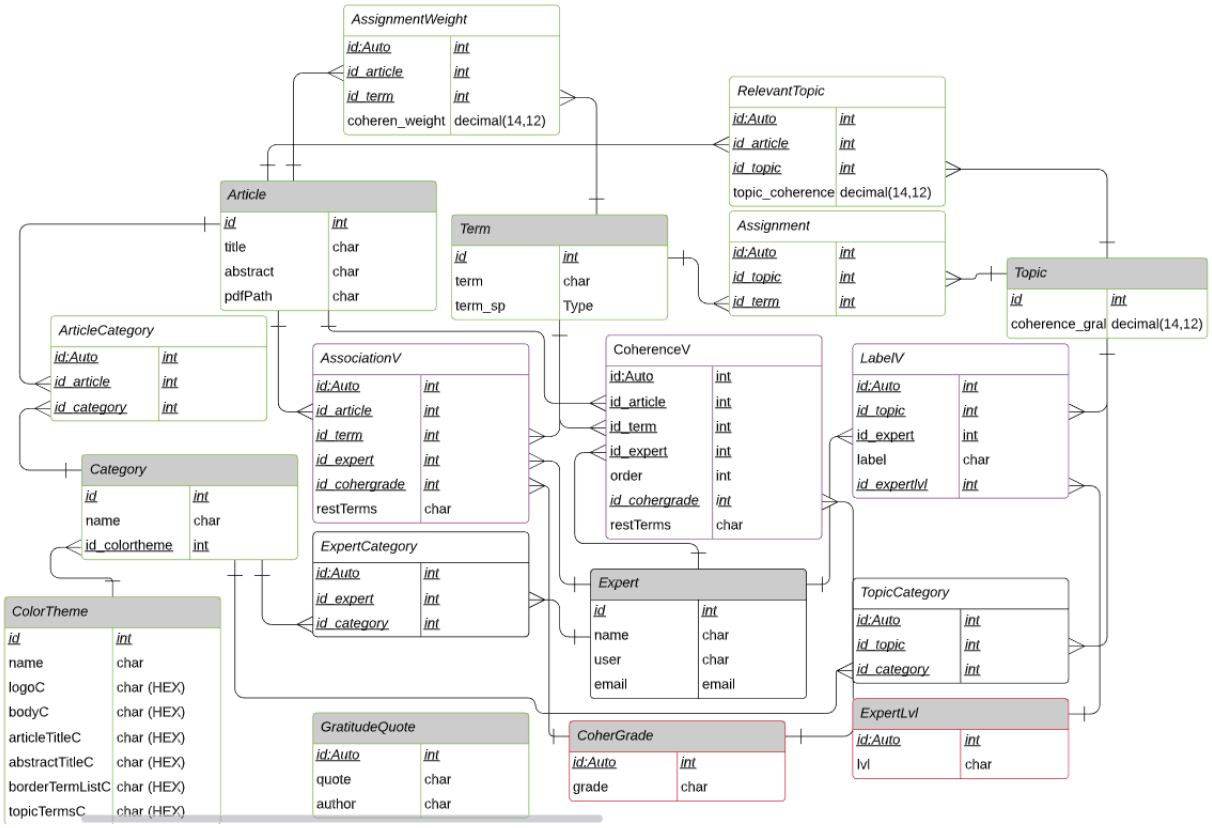


Figura F-11.: "whatTopic" v1.4.0. Modelo Entidad/Relación de la Base de datos. Elaboración propia.