eib0895841901

Algoritmo para el análisis temático de documentos digitales

Luis Roberto Polo Bautista*
Karen Vanessa Martínez Acevedo*

Artículo recibido:
21 de febrero de 2021
Artículo aceptado:
14 de junio de 2021

Artículo de investigación

Resumen

El objetivo del artículo es presentar un algoritmo para asignar áreas temáticas a documentos digitales que sirva como herramienta de apoyo al análisis temático dentro de la organización de la información, con el fin de ser implementado en el desarrollo de vocabularios controlados. La metodología utilizada consistió en aplicar el Reconocimiento Óptico de Caracteres (ROC) y la Asignación Latente de Dirichlet (ALD) como las principales herramientas para el desarrollo de un algoritmo basado en el lenguaje de programación Python, que permite la lectura de archivos con extensión PDF para la obtención de los principales temas del *corpus* textual. Los resultados de la aplicación del

* Escuela Nacional de Biblioteconomía y Archivonomía, México
luispolo221@yahoo.com.mx miwalakia83@gmail.com

algoritmo demuestran su utilidad en el área de la indización como un sistema para identificar y extraer temas relevantes de un documento específico en formato electrónico, permitiendo la automatización de procesos por parte del profesional de la información. De esta forma, se concluye su uso como desarrollo de puntos de acceso alternativos en función del contenido de los textos.

Palabras clave: Asignación Latente de Dirichlet; Algoritmos; Análisis Temático; Documentos Digitales

Algorithm for thematic analysis of digital documents

Luis Roberto Polo Bautista and *Karen Vanessa Martínez Acevedo*

Abstract

The objective of the article is to present an algorithm for assigning subject areas to digital documents which serve as a support tool for thematic analysis within the organization of information, in order to be implemented in development of controlled vocabularies. The methodology used consisted in applying Optical Character Recognition (OCR) and Latent Dirichlet Allocation (LDA) as main tools for developing an algorithm based on Python programming language,which allows reading of files with a PDF extension in order to obtain the main themes of textual corpus. Results of the algorithm's application demonstrate its usefulness in the area of indexing as a system for identifying and extracting relevant topics from a specific document in electronic format, and allow automation of processes by the information professional. This way, its use as a development of alternative points of access based on the content of texts is concluded.

Keywords: Latent Dirichlet Allocation; Algorithms; Thematic Analysis; Digital Documents

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2021.89.58419

Introducción

En la actualidad, gran parte del conocimiento colectivo se encuentra digitalizado y almacenado en bases de datos en forma de noticias, páginas de internet, literatura científica, entre otras formas, lo que dificulta el proceso de búsqueda de información por parte de un usuario específico (Blei, 2012: 77). Por lo tanto, una tendencia dentro del ámbito bibliotecológico referente a solucionar el problema anterior es la indización y recuperación temática tomando como base el estudio de la organización de contenidos en internet (Naumis, 2015: 223).

Para ello, "necesitamos nuevas herramientas computacionales para ayudar a organizar, buscar y comprender esta gran cantidad de información" (Blei, 2012: 77). Piepenbrink y Gaur (2017: 11335) mencionan que, a "pesar de muchos avances en el dominio del análisis de contenido, todavía carecemos de herramientas sofisticadas que se puedan utilizar para analizar grandes datos textuales […] como un enfoque del big data para identificar temas en el análisis de contenido de documentos digitales".

"Desde el punto de vista de los desarrolladores de tecnología, la solución está en disponer de herramientas potentes y sumamente rápidas, que nos ayuden a tomar decisiones proactivas, y conducidas por un conocimiento acabado de la información" (Naumis, 2015: 224). Los métodos que pueden utilizar estas herramientas pueden tener distintas finalidades, como la extracción de términos, la clasificación, la agrupación de documentos y la identificación de relaciones (Contreras, 2018: 111).

En las bibliotecas tradicionales, el proceso de indización es efectuado de forma manual e intelectual por personal especializado, dando como resultado palabras clave en forma de descriptores o encabezamientos de materia que son utilizados como puntos de acceso en la búsqueda de la información (Contreras, 2018: 112). De acuerdo con la norma UNE 50-121-91 emitida por la Asociación Española de Normalización y Certificación (1991: 3), el proceso de indización consiste en las siguientes etapas:

a) Examen del documento y determinación de su contenido: las partes más importantes del texto deben examinarse cuidadosamente y se debe prestar atención al título, resumen (si lo tiene), sumario o tabla de contenido, introducción, párrafos iniciales de capítulos y conclusiones.
b) Identificación y selección de los conceptos principales del contenido: el indizador debe identificar los términos más apropiados que reflejen las nociones esenciales de la descripción del contenido.

c) Selección de los términos de indización: el indizador debe verificar si los términos identificados en el paso anterior se encuentran en diccionarios y enciclopedias de autoridad, tesauros u clasificaciones temáticas.

Con base en estas etapas y al control de calidad respecto a la asignación temática mencionada en la misma norma, se podría mencionar que la indización tradicional es un proceso que conlleva tiempo y esfuerzo adicional por parte del bibliotecario profesional, considerando que en algunos casos particulares debe conocer o estar familiarizado con el área de conocimiento de un documento, con el objetivo de identificar y comprender con mayor precisión el vocabulario utilizado.

Asimismo, el proceso de indización puede conllevar a la asignación de descriptores o encabezamientos erróneos que no describan la totalidad o una parte esencial del contenido de un documento, dificultando el proceso de recuperación de información por parte de un usuario específico.

El algoritmo presentado en este artículo pretende optimizar de forma eficiente el tiempo de procesamiento de información, así como mejorar la identificación de temas de un documento, empleando el método de extracción de términos o palabras clave mediante el modelo de Asignación Latente de Dirichlet (ALD, en inglés Latent Dirichlet Allocation) por medio del aprendizaje no supervisado, así como también el Reconocimiento Óptico de Caracteres (ROC, en inglés Optical Character Recognition) para la conversión del texto completo de documentos digitales en archivos editables.

La ALD es un modelo probabilístico generativo de un *corpus*, cuya idea básica es que los documentos se representan como mezclas aleatorias sobre temas latentes, donde cada tema se caracteriza por una distribución sobre palabras. Es importante mencionar que el modelo no está necesariamente ligado al texto y tiene aplicaciones que involucran otros tipos de datos (Blei, Ng y Jordan, 2003: 996).

La notación y la terminología que utiliza el modelo ALD es la siguiente: una palabra es la unidad básica de datos discretos, definida como un elemento de un vocabulario indexado por {1,...,V}. Representamos palabras usando vectores de base unitaria que tienen un solo componente igual a uno y todos los demás componentes iguales a cero. Así, utilizando superíndices para denotar componentes, una palabra en una posición determinada en el vocabulario está representada por un V-vector w, de tal manera que $w^v = 1$ *y* $w^u = 0$ *para* $u \neq v$ (Blei, Ng y Jordan, 2003: 995).

"Un documento es una secuencia de N palabras indicadas por $w = (w_1, w_2,...,w_N)$ donde w_n es la posición de una palabra en la secuencia" (Blei, Ng y Jordan, 2003: 995).

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2021.89.58419

"Un corpus es una colección de M documentos indicados por $D = \{w_1, w_2, ...w_M\}$ " (Blei, Ng y Jordan, 2003: 995).

Por otro lado, el ROC es definido como "un software que convierte el texto impreso y las imágenes en forma digitalizada para que pueda ser manipulado por una máquina" (Islam, Islam y Noor, 2017: 1). A diferencia del cerebro humano, las computadoras no son lo suficientemente inteligentes para percibir la información disponible en una imagen o un documento impreso. Por lo tanto, se han presentado una gran cantidad de esfuerzos de investigación que intentan transformar este tipo de documentos a un archivo en un formato comprensible para una computadora (Islam, Islam y Noor, 2017).

El ROC asegura que la información de dichos documentos se digitalice a través de sistemas de tecnología de la información (TI) para su análisis. El procesamiento de lenguaje natural enriquece este proceso al permitir que esos sistemas reconozcan conceptos relevantes en el texto resultante, lo que es beneficioso para los análisis de aprendizaje automático (Smolaks, 2019). De esta manera, la tecnología ROC permitió convertir el texto completo de diversos documentos digitales en archivos editables con la finalidad de ser procesado por el modelo ALD. La combinación del ROC y el modelo ALD permitió la elaboración de un algoritmo basado en aprendizaje no supervisado que tome un documento en formato PDF como dato de entrada, convierta el texto completo a un archivo editable y automáticamente lo analice con el objetivo de mostrar los temas relevantes de ese documento de una forma dinámica.

El desarrollo de esta herramienta tiene como premisa principal facilitar la organización de la información por medio del análisis temático de documentos digitales, con la finalidad de ser implementado en la construcción de vocabularios controlados, como tesauros, taxonomías u ontologías que representen puntos de acceso alternativos en función al contenido de un texto.

Metodología

Como se mencionó anteriormente, el algoritmo se elaboró tomando como base el ROC y la ALD a través del lenguaje de programación Python por medio del entorno Jupyter Notebook, el cual es "una aplicación web de código abierto que permite crear y compartir documentos que contienen código, ecuaciones, visualizaciones y texto narrativo" (Project Jupyter, 2021).

El método empleado para la aplicación del algoritmo se divide en dos partes fundamentales (*Figura 1*). La primera se refiere a la conversión del texto completo del documento y la segunda corresponde a la identificación de temas y su modelación visual por medio de HTML. Cabe señalar que el proceso

para comprobar la similitud de los temas generados por el algoritmo y temas propios de un documento (palabras clave) se realizó a partir del índice de Jaccard, el cual se mostrará posteriormente.

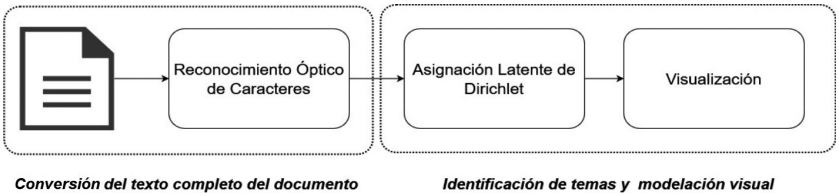

Figura 1. Proceso empleado para la aplicación del algoritmo

Conversión del texto completo del documento

En esta parte se utilizó el módulo PyMuPDF de Python, que entre otras funciones permite la identificación y conversión del texto de diferentes tipos de documentos a archivos editables. Es importante señalar que el funcionamiento de este algoritmo toma como base documentos en formato PDF, con el módulo PyMuPDF se hace uso del ROC, que permita generar archivos de texto simple (archivos con extensión txt) que contengan el texto completo de los documentos. A continuación, se muestra el pseudocódigo que se utilizó en este proceso:

Algoritmo: Conversión del texto completo de documentos
INICIO 1 Importar los módulos necesarios 2 Importar el documento en formato PDF 3 Leer el documento mediante el módulo PyMuPDF 4 Mostrar el número de páginas y los datos bibliográficos 5 Declarar una variable que contenga el nombre del nuevo archivo en formato txt 6 Para todas las páginas del documento hacer 7 Conversión del texto en formato UTF-8 8 Escribir el texto en un archivo con el formato definido en la línea 5 9 Definir dentro del archivo las divisiones de cada página 10 Cerrar el documento FIN

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2021.89.58419

Identificación de temas y modelación visual

En la sección de identificación de temas se utilizaron los módulos Numpy y Scikit-Learn para el cálculo de frecuencia y vectores de palabras y el modelo de Asignación Latente de Dirichlet (ALD) para la identificación de los temas relevantes. Para la visualización dinámica de los temas por medio de HTML se utilizó el módulo pyLDAvis. A continuación, se muestra el pseudocódigo que se utilizó en este proceso:

Algoritmo: Identificación de temas relevantes y modelación visual
INICIO 1 Importar los módulos necesarios 2 Importar el archivo de texto simple (archivo en formato txt) 3 Eliminar los artículos, pronombres, preposiciones, etc., del texto y normalizar el texto en letras minúsculas 4 Normalizar todas las formas de una misma palabra 5 Calcular los vectores de las palabras individuales que conforman cada frase 6 Crear una función que extraiga los vectores de las palabras y los guarde en una variable 7 Declarar en una variable el número de temas a procesar 8 Declarar en una variable el número de palabras relacionadas a los temas 9 Procesar los datos ingresados en la línea 6 con base al módulo ALD 10 Mostrar los datos procesados referentes a la línea 8 con base al vector asignado a cada palabra, así como el número de palabras relacionados a los temas 11 Declarar en una variable el nombre del archivo en formato HTML 12 Procesar los datos referentes a la línea 9 para que se puedan mostrar de forma dinámica 13 Preparar el archivo con los datos a mostrar 14 Guardar el archivo en una ruta determinada con extensión HTML FIN

INVESTIGACIÓN BIBLIOTECOLÓGICA, vol. 35, núm. 89, octubre/diciembre, 2021, México, ISSN: 2448-8321, pp. 13-31

Para analizar los temas obtenidos automáticamente mediante el algoritmo, existen diversos métodos que calculan la consistencia de la indización entre dos sistemas (manual y automatizado); en este trabajo se utilizó el índice de Jaccard para identificar el coeficiente de similitud entre los conjuntos de términos generados automáticamente y los conjuntos de términos generados manualmente de un documento específico.

Kosub (2019: 36) menciona que "el índice de Jaccard es una medición clásica de similitud con varias aplicaciones prácticas en la recuperación de información, extracción de datos, aprendizaje automático, entre otras". Campos (2017) describe el índice de Jaccard como "la división entre el número de elementos en común que tienen los dos conjuntos sobre el número de elementos únicos que tiene la unión de ambos conjuntos".

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

Donde:
A = Conjunto de datos 1
B = Conjunto de datos 2
$|A \cap B|$ = Número de elementos en común de ambos conjuntos
$|A \cup B|$ = Número de elementos únicos de ambos conjuntos

El algoritmo de análisis de temas descrito en la metodología fue implementado de forma individual a cada uno de los documentos de un conjunto de artículos científicos tomados de la base de datos SciELO, de tal manera que se pueda determinar el coeficiente de similitud por medio del índice de Jaccard entre los temas identificados automáticamente y los identificados de forma manual, que son principalmente palabras clave generadas por los propios autores.

Los artículos utilizados para este análisis fueron 50, divididos en cinco áreas del conocimiento (Ciencias agrarias, Ciencias biológicas, Ciencias de la salud, Ciencias exactas y de la tierra, y Ciencias sociales aplicadas) con 10 documentos de cada área; cinco en español y cinco en inglés.

A continuación, se muestra la aplicación del algoritmo en un artículo en español, donde se visualizan los resultados obtenidos. La *Figura 2* muestra un fragmento de texto del documento original y su modificación tras la eliminación de los artículos, pronombres, preposiciones, etc., así como su

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2021.89.58419

normalización a letras minúsculas (este proceso se encuentra descrito en la identificación de temas y modelación visual).

Este proceso de eliminación de los artículos, pronombres, preposiciones, entre otros, facilitó el tratamiento automático del documento, permitiendo una mayor precisión en los resultados referentes a la identificación de palabras clave como temas relevantes del *corpus* textual.

De acuerdo con Gil Leiva (1997: 120), la supresión de estos términos, que comúnmente se conocen como palabras vacías o *stopwords* dentro del ámbito de la bibliotecología y ciencias de la información, queda justificado por los siguientes motivos:

a) El descarte de las palabras vacías provoca que disminuya el número de palabras a procesar.
b) Se reduce en un menor tiempo el análisis. Uno de los objetivos generales que se persigue al automatizar la indización es que el tiempo empleado por el programa sea similar o inferior al de un profesional.
c) Es una ventaja no contar con este tipo de palabras en la etapa de búsqueda de términos construidos de forma diferente respecto a los términos autorizados.
d) "Para los textos de diferentes áreas temáticas e idiomas se ha comprobado que, aproximadamente el cincuenta por ciento de las palabras manejadas son palabras de este tipo".

Estas palabras se eliminaron utilizando el módulo NLTK de Python por medio de las *stopwords* en idioma español, así como también de una lista donde se incorporaron algunas palabras vacías que no contempla el módulo, con el objetivo de extender aquellas palabras que no se requieran considerar para el análisis.

En este artículo se afrontan los desafíos para acceder y tratar los fondos de las hemerotecas nacionales de Colombia, Ecuador, México y Uruguay, que recogen noticias sobre eventos meteorológicos entre los siglos XIX-XX. Sobre estos periódicos se conforma un corpus de noticias que mediante lecturas técnicas y la aplicación de un proceso de bibliominería, utilizando diversas herramientas, permite iniciar la construcción de una red de ontologías.	artículo afrontan desafíos acceder tratar fondos hemerotecas nacionales colombia ecuador méxico uruguay recogen noticias eventos meteorológicos siglos xixxx periódicos conforma corpus noticias lecturas técnicas aplicación proceso bibliominería herramientas iniciar construcción red ontologías

Figura 2. Texto original y texto modificado tras la eliminación de las *stopwords*

En la *Tabla 1* se muestran los cuatro principales temas identificados por el algoritmo, considerando su frecuencia general.

Palabra	Frecuencia general
Noticias	34
Ontologías	29
Corpus	27
Meteorológicos	25

Tabla 1. Temas relevantes

En el trabajo de Chuang, Manning y Heer (2012: 75) se menciona que la frecuencia general de las palabras se basa en el cálculo de la prominencia del término, tal como se describe a continuación: para una palabra dada *w*, se calcula su probabilidad condicional *P* (*T*|*w*): la probabilidad de que la palabra observada *w* fue generada por tema latente *T*. También se calcula la probabilidad marginal *P*(*T*): la probabilidad de que cualquier palabra *w´* fue generada por tema *T* . Se define la distinción de la palabra como la divergencia Kullback-Leibler entre *P* (*T*|*w*) y *P*(*T*):

$$Diferencia\ (w) = \sum_T P(T|w)\ log\ \frac{P(T|w)}{P(T)}$$

De igual manera, Chuang, Manning y Heer (2012: 75) mencionan que esta formulación describe (en un sentido teórico de la información) cuán informativo es el término específico *w* para determinar el tema generador, en comparación con un término seleccionado al azar *w´*. Por ejemplo, si una palabra *w* ocurre en todos los temas, la observación de la palabra nos dice poco sobre la mezcla temática del documento; por lo tanto, la palabra recibiría una puntuación de distinción baja. La calidad de un término está definida por el producto:

$$Prominencia(w) = P(w)\ x\ Diferencia\ (w)$$

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2021.89.58419

En la *Tabla 2* se muestra una comparación entre los temas identificados por el algoritmo y las palabras clave propuestas por el autor del artículo, con la finalidad de calcular su coeficiente de similitud. Como se puede observar, existen palabras entre ambos conjuntos de términos que son semánticamente similares, como: *Meteorología – Meteorológicos; Ontología – Ontologías; Prensa – Noticias.*

Por el otro lado, la palabra *Bibliominería* no tiene un término similar en los temas obtenidos a través del algoritmo, de lo que se infiere que esa palabra hace referencia a una metodología aplicada en el artículo y su frecuencia y relevancia son mínimas en comparación con otras.

El resultado del coeficiente de similitud con base al índice de Jaccard demuestra que existe un alto porcentaje de semejanza entre ambos conjuntos de temas, lo que indica que el grado de precisión de identificación y extracción de temas del algoritmo es adecuado.

Palabras clave propuestas por el autor	**Temas identificados por el algoritmo**
Bibliominería	Corpus
Meteorología	Meteorológicos
Ontología	Ontologías
Prensa	Noticias
Coeficiente de similitud	
J (A, B) = 3 / 4 = 0.75 =75 %	

Tabla 2. Cálculo del coeficiente de similitud

En la *Figura 3* se presenta la visualización de la modelación de temas en forma dinámica por medio de HTML. Esta representación visual fue elaborada mediante el módulo pyLDAvis y tiene la finalidad de responder algunas preguntas básicas sobre el modelado de temas, tales como ¿cuál es el significado de cada tema?, ¿qué tan prevalente es cada tema? y ¿cómo se relacionan los temas entre sí? Distintos componentes visuales responden a cada pregunta, de los cuales algunas herramientas son originales y otras son adaptadas de modelos ya existentes (Sievert y Shirley, 2014: 63).

El panel izquierdo presenta una vista global de los temas y responde las dos últimas preguntas. En esta vista, se trazan los temas como círculos en el plano bidimensional cuyos centros se determinan al calcular la distancia entre temas, y luego mediante el uso de escalas multidimensionales para proyectar las distancias entre temas en dos dimensiones (Sievert y Shirley, 2014: 63).

El panel derecho muestra una gráfica de barras horizontal, que representa los términos individuales que son más útiles para interpretar el tema

actualmente seleccionado a la izquierda, y permite a los usuarios responder la primera pregunta. Un par de barras superpuestas representan tanto la frecuencia de un término dado como la frecuencia específica de un tema correspondiente con un término (Sievert y Shirley, 2014: 63). Esta superposición se puede visualizar en el archivo HTML original.

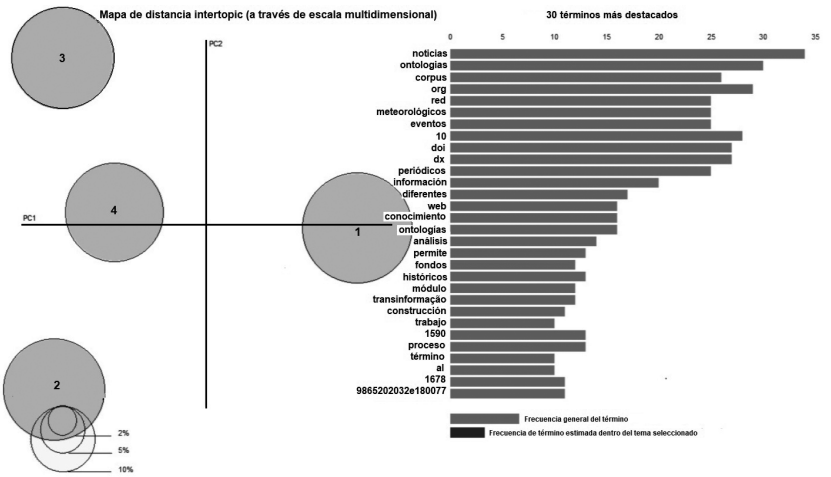

Figura 3. Visualización dinámica de temas

Las limitaciones que tiene esta visualización son que en algunos casos contempla palabras que originalmente debería omitir, como artículos, números o abreviaciones, aun cuando en el algoritmo se eliminan. Sin embargo, no representan un problema para la mayoría de los temas identificados, los cuales son los que se deberían de tomar en cuenta.

En la *Figura 4* se presentan los porcentajes del coeficiente de similitud obtenidos mediante el índice de Jaccard, aplicado a cada uno de los 50 artículos mencionados. La distribución es útil para representar las características de la dinámica de crecimiento y descenso entre los coeficientes, de tal manera que si el porcentaje se acerca a 100 hace referencia a que el total de palabras clave generadas por los autores y los temas identificados automáticamente son similares; por el contrario, si el porcentaje se acerca a 0 estos son diferentes entre sí. De acuerdo con lo mostrado en la gráfica, podemos mencionar que el promedio de similitud de todos los artículos analizados es de ≈ 69 %.

El promedio general del coeficiente de similitud tras la aplicación del algoritmo en los artículos en español fue de ≈ 66 %, el promedio en los artículos

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2021.89.58419

en inglés fue de ≈ 72 %. Con base en esto, podemos mencionar que, aunque el promedio de similitud fue más alto en los artículos en inglés, no indica que el rendimiento del algoritmo fue mejor, ya que existen diversos factores que pueden generar variaciones en los promedios de similitud, como la diferencia entre las palabras clave y los temas obtenidos.

En algunos casos, las palabras clave generadas por los autores estaban constituidas por una cadena de términos, esto causó un inconveniente al realizar la comparación, ya que el algoritmo sólo identifica temas conformados por una palabra (token). Para ello, se consideraron aquellas palabras claves semánticamente similares a los temas obtenidos automáticamente, y viceversa.

Para calcular la eficiencia del algoritmo, se tomó en cuenta el tiempo de ejecución considerando la extensión del documento. La ejecución del algoritmo se probó en una computadora con las siguientes especificaciones básicas: procesador Intel (R) Celeron (R) 1.60 GHz, memoria RAM de 4 GB, sistema operativo de 64 bits y un procesador x64. Para calcular el tiempo de ejecución se utilizaron cuatro documentos en formato PDF con diferente extensión (cantidad de hojas), con la finalidad de medir el tiempo de procesamiento de la información con base en el análisis temático. En las *Figuras 5* y *6* se muestran los resultados obtenidos.

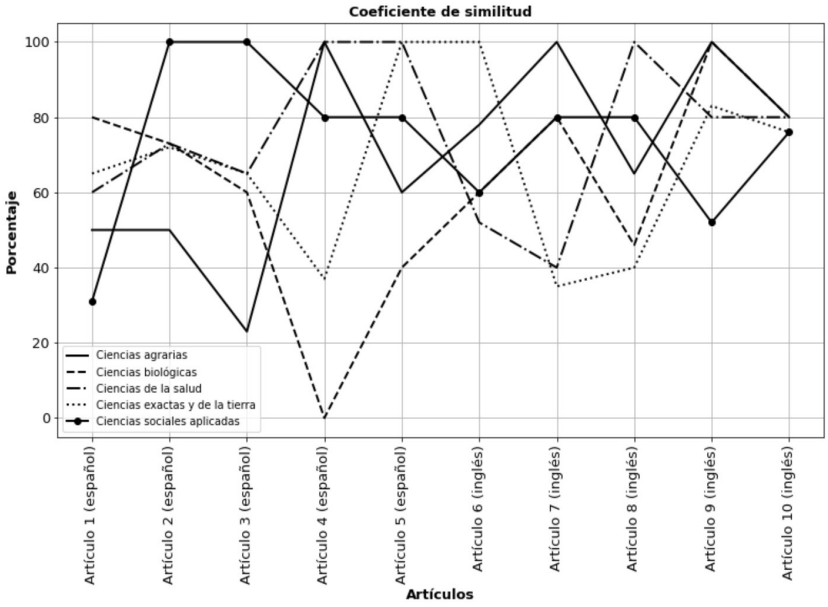

Figura 4. Coeficiente de similitud del corpus de artículos científicos

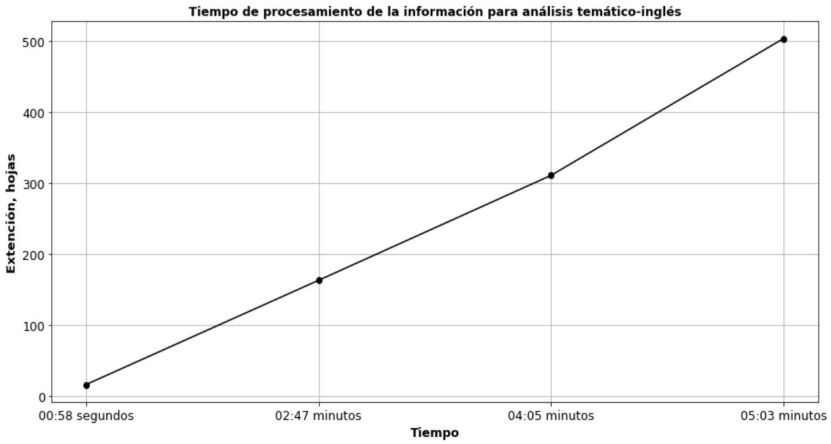

Figura 5. Tiempo de procesamiento de la información para análisis temático-inglés

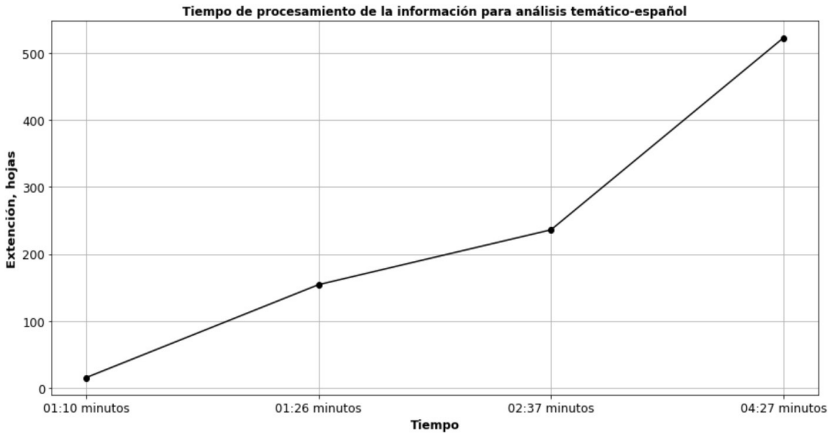

Figura 6. Tiempo de procesamiento de la información para análisis temático-español

Como se puede observar, el tiempo de ejecución del algoritmo referente al procesamiento de la información para el análisis temático de documentos es directamente proporcional a la extensión del documento, lo que indica que al aumentar una variable, en este caso la extensión del documento, también aumenta el tiempo de procesamiento.

De esta forma, podemos mencionar que el promedio de tiempo para procesar un documento con una extensión de entre 10-16 hojas y en idioma inglés o español es de 01:24 minutos y, en el otro extremo, el promedio de tiempo para procesar un documento con una extensión de 500-550 hojas en idioma inglés o español es de 05:05 minutos.

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2021.89.58419

De acuerdo con Delgado y Sosa (1998), la obtención de un libro procesado completamente como producto final en las bibliotecas tradicionales consume en promedio, aproximadamente, más de una hora y media. Los tiempos promedios por cada tarea son los siguientes: catalogar: 24:09 minutos, clasificar e indizar: 46:08 minutos, preparación física del documento: 06:06 minutos, y la creación de fichas bibliográficas: 18:06 minutos.

Con base en lo anterior, consideramos que la utilización de este algoritmo como apoyo en el análisis temático de documentos digitales (indización) podría reducir significativamente el tiempo estimado para la tarea de clasificar e indizar, a aproximadamente menos de la mitad de tiempo.

Discusión

Algunos inconvenientes al hacer uso de vocabularios controlados tradicionales como herramientas para el análisis temático de la información son los siguientes:

a) El control de las formas plural y singular de las palabras por medio de la sinonimia y polisemia puede generar consecuencias negativas en la búsqueda y recuperación de la información, principalmente por la agrupación de diversos términos unificándolos en uno solo.
b) Algunos encabezamientos o descriptores no reflejan el contenido de un documento específico.
c) El tiempo que conlleva el proceso de indización es mayor.
d) La cantidad de información a considerar para la asignación de temas es poca, y a veces no es suficiente para asignar correctamente un descriptor o encabezamiento.
e) El profesional de la información debe estar familiarizado con el área de conocimiento del documento o comprender los conceptos que se encuentren en su contenido.

Por el contrario, las principales ventajas de la utilización de este algoritmo en el análisis temático son:

a) Optimización de tiempo por parte del bibliotecario profesional.
b) Mayor cantidad de información para procesar y en menor tiempo.
c) Mayor precisión en la identificación de temas relevantes de un documento.
d) Aunque en este trabajo sólo implementamos el algoritmo en idioma español e inglés, también se puede utilizar para documentos con diferentes idiomas.

e) La implementación de este algoritmo en las bibliotecas no contempla costos financieros altos.
f) Las habilidades tecnológicas que requiere el bibliotecario profesional para hacer uso de este algoritmo permiten una mejor adaptabilidad para manejar y gestionar diversas herramientas que puedan automatizar procesos a gran escala dentro de la biblioteca.

La identificación de temas (modelación de tópicos) generalmente se relaciona con las ciencias de la computación, específicamente con el aprendizaje automático y el procesamiento de lenguaje natural; sin embargo, en la actualidad, este tipo de herramientas tecnológicas pueden ser aprovechadas dentro de la bibliotecología y ciencias de la información como métodos de automatización de procesos. La identificación de temas puede ser implementada dentro del proceso de análisis temático en la organización de la información, con la finalidad de identificar temas relevantes en un documento específico.

En varias disciplinas se han desarrollado métodos que utilizan la identificación temática de un *corpus* de datos, generalmente asociado a textos cortos y minería de texto. De igual manera, dentro de la bibliotecología se han desarrollado varios procedimientos similares que tienen como objetivo automatizar el proceso de indización dentro de las unidades de información.

La principal diferenciación del algoritmo descrito en este artículo con los métodos anteriormente desarrollados recae en la utilización del ROC como herramienta de conversión de un documento a un archivo editable que contenga el texto completo de un documento, con la finalidad de facilitar el tratamiento y análisis de la información, así como también la optimización de tiempo y procesos por parte del profesional de la información.

Si bien en algunos casos particulares el rendimiento del algoritmo se vio afectado por la inclusión de palabras vacías como abreviaciones, pronombres, artículos, tal como se mostró en la *Figura 3*, estos no representan mayor inconveniente para seleccionar aquellos temas que reflejen el contenido del documento. De igual manera, un inconveniente del algoritmo es que sólo identifica temas compuestos por una única palabra (token), y generalmente existen áreas temáticas que deben expresarse con dos o más palabras.

Es importante señalar que, en los casos en que el porcentaje de similitud fue bajo en un documento, no significa que los temas identificados automáticamente no hayan sido adecuados, sino que las palabras clave propuestas por los autores generalmente no reflejaban la totalidad del contenido del artículo o su relevancia dentro del *corpus* textual era mínima.

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2021.89.58419

Conclusiones

Presentamos un algoritmo que permite identificar temas relevantes de un documento digital en formato PDF, utilizando el Reconocimiento Óptico de Caracteres y la Asignación Latente de Dirichlet a través del aprendizaje no supervisado, lo que indica que el algoritmo no necesita datos de entrenamiento para poder generar temas basados en el contenido del documento.

La utilización de este algoritmo identifica ≈ 30 temas relevantes de un documento específico en idioma español o inglés, dando oportunidad de seleccionar aquellos que mejor representan el contenido del recurso bibliográfico.

La implementación del algoritmo permite procesar una gran cantidad de información en un tiempo relativamente menor en comparación con un bibliotecario profesional, generando temas que reflejen la totalidad del contenido de un documento.

El promedio del índice de similitud entre los temas generados automáticamente y los propuestos por los autores (palabras clave) del *corpus* de artículos científicos utilizado es de ≈ 69 %, lo que indica que más de la mitad de los temas generados por el algoritmo fueron semejantes a las palabras clave propuestas por los autores.

Como se mencionó anteriormente, el algoritmo calcula los temas relevantes de un documento con base a una distribución de las palabras que contiene, de esta forma no afectaría de qué área de conocimiento tratase el documento, la información del contenido se procesaría de igual forma teniendo relativamente los mismos resultados de tiempo de ejecución.

La utilización de este algoritmo en el proceso de análisis temático puede solucionar algunos inconvenientes como la optimización del tiempo, la cantidad de información a procesar, la asignación de temas *ad hoc* y la compatibilidad con el idioma del documento.

La implementación de este algoritmo dentro de las bibliotecas no implica costos financieros altos, ya que esta herramienta se puede replicar en una computadora con especificaciones básicas, y al utilizar este tipo de sistemas permite que el bibliotecario profesional obtenga nuevos conocimientos y habilidades que permitan un mejor desempeño en otras áreas, así como mejorar el aprovechamiento de tiempo.

Este algoritmo se puede utilizar para el desarrollo de puntos de acceso alternativos en función al contenido de los textos, así como la creación de vocabularios controlados, como tesauros, taxonomías u ontologías, que representen un dominio de conocimiento específico mediante el análisis del discurso.

Cabe mencionar que este algoritmo puede tener algunas mejoras potenciales relacionadas a la identificación de temas y su estructura en más de una

palabra (token), así como aumentar la cantidad de idiomas que pueda permitir en un documento.

En términos generales, los bibliotecarios profesionales pueden aprovechar este tipo de herramientas tecnológicas para crear sistemas capaces de automatizar procesos dentro de una biblioteca, como catalogar, indizar, clasificar, entre muchos otros, con la finalidad de darle una nueva perspectiva a la biblioteca y por ende a la bibliotecología y ciencias de la información.

Referencias

Asociación Española de Normalización y Certificación. 1991. *Métodos para el análisis de documentos, determinación de su contenido y selección de los términos de indización*. https://nanopdf.com/download/metodos-para-el-analisis-de-documentos-determinacion-de-su_pdf

Blei, David. 2012. "Probabilistic topic models". *Communications of the ACM* 55 (4): 77-84. https://doi.org/10.1145/2133806.2133826

Blei, David, Andrew Ng y Michael Jordan. 2003. "Latent Dirichlet Allocation". *Journal of Machine Learning Research* (3): 993-1022. https://web.archive.org/web/20120207011313/http://jmlr.csail.mit.edu/papers/volume3/blei03a/blei03a.pdf

Campos, Diego. 2017. "Métricas de similitud para cadenas de texto. Parte III". *SoldAI*, 24 de diciembre de 2019. https://blog.soldai.com/metricas-conjuntos-emparejamiento-caracteres/

Chuang, Jason, Christopher Manning y Jeffrey Heer. 2012. "Termite: Visualization Techniques for Assessing Textual Topic Models". Trabajo presentado en la 12th International Working Conference on Advanced Visual Interfaces, Capri Island, Italy, 21-25 de mayo. http://vis.stanford.edu/files/2012-Termite-AVI.pdf

Contreras, Marcial. 2018. "Aplicación del algoritmo RAKE en la indización de documentos digitales". *Investigación Bibliotecológica: archivonomía, bibliotecología e información* 32 (75): 109-123. https://doi.org/10.22201/iibi.24488321xe.2018.75.57951

Delgado, Nora e Hilda Sosa. 1998. "Evaluación de la eficiencia en bibliotecas". *Investigación bibliotecológica: archivonomía, bibliotecología e información* 12 (24): 57-82.

Gil Leiva, Isidoro. 1997. "La automatización de la indización: propuesta teórica-metodológica. Aplicación en el área de biblioteconomía y documentación". Tesis de licenciatura, Universidad de Murcia, Departamento de Información y Documentación, España. http://hdl.handle.net/10803/10917

Islam, Noman, Zeeshan Islam y Nazia Noor. 2017. "A survey on optical character recognition system". *Journal of Information & Communication Technology* 10 (2): 1-4. https://arxiv.org/ftp/arxiv/papers/1710/1710.05703.pdf

Kosub, Sven. 2019. "A note on the triangle inequality for the Jaccard distance". *Pattern Recognition Letters* 120, 36-38. https://doi.org/10.1016/j.patrec.2018.12.007

Naumis, Catalina. 2015. "Tendencias en la indización y recuperación temática", en *La información: perspectivas bibliotecológicas y distinciones interdisciplinarias*, coordinado por Jaime Ríos Ortega y César Augusto Ramírez Vélazquez, 223-240. México: UNAM, Instituto de Investigaciones Bibliotecológicas y de la Información. http://ru.iibi.unam.mx/jspui/bitstream/IIBI_UNAM/L105/1/informacion_perspectivas_bibliotecologicas.pdf

Piepenbrink, Anke y Ajai Gaur. 2017. "Topic models as a novel approach to identify themes in content analysis". *Academy of Management Proceedings* 2017 (1): 11335. https://doi.org/10.5465/ambpp.2017.141

Project Jupyter. 2021. *The Jupyter Notebook*. https://jupyter.org/

Sievert, Carson y Kenneth Shirley. 2014. "LDAvis: A method for visualizing and interpreting topics". Trabajo presentado en Workshop on Interactive Language Learning, Visualization, and Interfaces, Baltimore, Maryland, USA, 27 de junio. https://doi.org/10.3115/v1/w14-3110

Smolaks, Max. 2019. "Intelligent document imaging with natural language processing and optical character recognition". *AI Business*, 16 de agosto de 2019. https://aibusiness.com/document.asp?doc_id=761011

Para citar este texto:

Polo Bautista, Luis Roberto y Karen Vanessa Martínez Acevedo. 2021. "Algoritmo para el análisis temático de documentos digitales". *Investigación Bibliotecológica: archivonomía, bibliotecología e información* 35 (89): 13-31. http://dx.doi.org/10.22201/iibi.24488321xe.2021.89.58419

Anexo. Visualización y descarga del algoritmo

El algoritmo de análisis temático de documentos digitales presentado en este artículo se puede visualizar y descargar completo por medio del siguiente enlace: https://github.com/LuisPoloBautista/Algoritmo_para_el_analisis_tematico_de_documentos_digitales

Antes de ejecutar el algoritmo, asegúrese de tener en cuenta las recomendaciones mencionadas en el archivo "README".

DOI: http://dx.doi.org/10.22201/iibi.24488321xe.2021.89.58419