Universidad Nacional Mayor de San Marcos
Universidad del Perú. Decana de América
Facultad de Letras y Ciencias Humanas
Escuela Profesional de Bibliotecología y Ciencias de la Información

El etiquetado social en la descripción de libros: una comparación entre etiquetas y encabezamientos de materia

TESIS

Para optar el Título Profesional de Licenciada en Bibliotecología y Ciencias de la Información

AUTOR

Betty Nancy LOAYZA PALACIOS

ASESOR

Carlos Javier ROJAS LAZARO

Lima, Perú

2022

Referencia bibliográfica

Loayza, B. (2022). *El etiquetado social en la descripción de libros: una comparación entre etiquetas y encabezamientos de materia*. [Tesis de pregrado, Universidad Nacional Mayor de San Marcos, Facultad de Letras y Ciencias Humanas, Escuela Profesional de Bibliotecología y Ciencias de la Información]. Repositorio institucional Cybertesis UNMSM.

Metadatos complementarios

Datos de autor	
Nombres y apellidos	Betty Nancy Loayza Palacios
Tipo de documento de identidad	DNI
Número de documento de identidad	43840696
URL de ORCID	https://orcid.org/0000-0002-8625-9966
Datos de asesor	
Nombres y apellidos	Carlos Javier Rojas Lazaro
Tipo de documento de identidad	DNI
Número de documento de identidad	08954040
URL de ORCID	https://orcid.org/0000-0001-8291-8138
Datos del jurado	
Presidente del jurado	
Nombres y apellidos	Martín Alonso Estrada Cuzcano
Tipo de documento de identidad	DNI
Número de documento de identidad	08435943
Miembro del jurado 1	
Nombres y apellidos	Carlos Antonio Sam Anlas
Tipo de documento de identidad	DNI
Número de documento de identidad	40789757
Miembro del jurado 2	
Nombres y apellidos	Filiberto Felipe Martínez Arellano
Tipo de documento de identidad	Pasaporte
Número de documento de identidad	MX / G25985897
Datos de investigación	
Línea de investigación	E.2.3.3. Tecnologías de la información y desarrollo de la investigación académica
Grupo de investigación	No aplica.

Agencia de financiamiento	Sin financiamiento.
Ubicación geográfica de la investigación	País: Perú Departamento: Lima Provincia: Lima Distrito: Lima Latitud: -12.05703 Longitud: -77.08154
Año o rango de años en que se realizó la investigación	2021-2022
URL de disciplinas OCDE	Bibliotecología https://purl.org/pe-repo/ocde/ford#5.08.03 Ciencias de la Información https://purl.org/pe-repo/ocde/ford#5.08.02

UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS
FACULTAD DE LETRAS Y CIENCIAS HUMANAS

ESCUELA PROFESIONAL DE BIBLIOTECOLOGÍA Y CIENCIAS DE LA INFORMACIÓN

ACTA DE SUSTENTACIÓN DE TESIS

A los veinticuatro días del mes de mayo del dos mil veintidós, a las dieciséis horas, en acto público se conecta por vía remota el Jurado de sustentación integrado por los siguientes profesores del Departamento Académico de Bibliotecología y Ciencias de la Información de la Facultad de Letras y Ciencias Humanas de la Universidad Nacional Mayor de San Marcos:

Martin Alonso Estrada Cuzcano	Presidente
Carlos Javier Rojas Lazaro	Asesor
Carlos Antonio Sam Anlas	Miembro
Filiberto Felipe Martínez Arellano	Experto externo

Con el fin de recibir la sustentación de Tesis: ***EL ETIQUETADO SOCIAL EN LA DESCRIPCIÓN DE LIBROS: UNA COMPARACIÓN ENTRE ETIQUETAS Y ENCABEZAMIENTOS DE MATERIA,*** presentada por la bachiller **BETTY NANCY LOAYZA PALACIOS.** Concluida la sustentación, el jurado procedió a la calificación con el siguiente resultado:

Aprobado con máximos honores

Números (19) Letras (diecinueve)

Luego del proceso de sustentación y la calificación correspondiente, se le comunicó al graduando el resultado obtenido y el Jurado recomienda a la Facultad que se le otorgue el título profesional de **Licenciada** en Bibliotecología y Ciencias de la Información.

Siendo las diecisiete horas, se concluyó el acto por lo cual, los miembros del jurado dan fe de lo actuado firmando la presente Acta.

Dr. Martin Alonso Estrada Cuzcano
Presidente (Principal)

Mg. Carlos Rojas Lazaro
Asesor (Auxiliar)

Mg. Carlos Antonio Sam Anlas
Miembro (Auxiliar)

Dr. Filiberto Felipe Martínez Arellano
Experto externo

Índice de contenido

Índice de tablas

Índice de figuras

Resumen

Como resultado de la representación del contenido de un recurso por parte de los usuarios de comunidades virtuales de catalogación, el etiquetado social guarda semejanza con el proceso de análisis temático de recursos de información en bibliotecas. El objetivo de la presente investigación fue determinar el grado de similitud entre los términos asignados por los usuarios de LibraryThing y los catalogadores de la Library of Congress para describir el contenido temático de libros de ficción y el de aquellos de no ficción. La recolección de datos hizo uso de la minería de textos, mientras que el análisis estadístico tuvo lugar a través del cálculo de coeficientes de similitud y correlación. En ambos casos se empleó a R como herramienta de análisis de datos. Los resultados sugieren que las etiquetas sociales presentan coincidencias con los encabezamientos de materia de la Library of Congress pero son lo suficientemente distintas para considerarlas como un complemento en la descripción de contenidos.

Palabras clave: folksonomías, etiquetado colaborativo, etiquetado social, lenguaje natural, encabezamientos de materia

Línea de investigación: Tecnologías de la información y desarrollo de la investigación académica

Abstract

As a result of the description of resources by users of social cataloging applications, social tagging resembles the thematic analysis of information resources in libraries. The purpose of this study was to determine the degree of similarity between the terms assigned by LibraryThing users and Library of Congress catalogers to describe the thematic content of fiction books as well as that of nonfiction books. The data collection employed text mining techniques, while the statistical analysis entailed the calculation of similarity and correlation coefficients. In both cases R was used as a data analysis tool. The results suggest that social tags show similarities with Library of Congress subject headings but are distinct enough to be considered as a complement in terms of content description.

Keywords: folksonomies, collaborative tagging, social tagging, natural language, subject headings

Line of research: Information technologies and development of academic research

Introducción

La descripción de contenidos temáticos en comunidades virtuales de catalogación difiere de las prácticas adoptadas por archivos, bibliotecas y repositorios institucionales. A diferencia de estos, lo que suele primar no es el uso de reglas y vocabularios controlados, sino la interpretación que un usuario particular pueda darle al contenido de un recurso, la cual es plasmada a través de términos provenientes del lenguaje natural. Ahora bien, a medida que este proceso tenga lugar en un contexto social, el intercambio de información propiciará que ciertos términos prevalezcan sobre otros, generando vocabularios más estables.

Teniendo en cuenta que el etiquetado social puede llegar a tener coincidencias con el proceso de descripción del contenido en el ámbito bibliotecario, cabe considerarlo como objeto de estudio. Para ello se realizó una comparación entre las etiquetas asignadas por los usuarios de LibraryThing y los encabezamientos de materia asignados por los catalogadores de la Library of Congress (Biblioteca del Congreso de los Estados Unidos de América), recurriendo a la minería de textos y el análisis estadístico para medir la similitud entre ambos conjuntos de datos.

Se partió de la hipótesis, refrendada por estudios previos, de que las etiquetas sociales tienen coincidencias parciales con respecto a los encabezamientos de materia. Esto resultó ser cierto, permitiendo concluir que ambos vocabularios podrían complementarse entre sí, incrementando las posibilidades de mejorar la accesibilidad a los recursos de información.

Desglosando el contenido de la presente investigación, el primer capítulo comprende los antecedentes y el marco teórico. Se define el concepto de folksonomía y se enumeran algunas características de los sistemas de etiquetado social, así como las ventajas y desventajas de las etiquetas frente a los vocabularios controlados. Asimismo, se

mencionan ejemplos de su aplicación en bibliotecas juntamente con estrategias de mejora del etiquetado social.

El segundo capítulo presenta la hipótesis y la variable de estudio, considerando una escala de valores para interpretar el nivel de similitud entre las etiquetas sociales y los encabezamientos de materia. Mientras tanto, el tercer capítulo detalla los procedimientos para la recolección de datos a través del uso de R, el cual cumple las funciones de lenguaje de programación y entorno para computación estadística y gráficos.

El quinto capítulo gira alrededor de los resultados, haciendo la distinción entre los conjuntos de datos según su naturaleza (etiquetas sociales y encabezamientos de materia), idioma (inglés y español) y tipo de literatura de los títulos seleccionados para el análisis (ficción y no ficción). Por último, el sexto capítulo presenta las conclusiones y las recomendaciones de acuerdo con las hipótesis planteadas.

Capítulo I: Problema de investigación

Descripción del problema de investigación

Atribuida al desarrollo científico surgido a partir de la Segunda Guerra Mundial, la explosión informativa no solamente ha permanecido como una constante hasta nuestros días, sino que se ha visto incrementada aún más gracias a la masificación de las tecnologías de la información y la comunicación. Si antes se requería de amplias bóvedas para albergar el conocimiento escrito en existencia, actualmente el equivalente a cientos de volúmenes físicos puede ser almacenado en unos cuantos gigabytes y puesto en línea para su libre acceso desde cualquier parte del mundo. Ello, sumado a la democratización en la producción de contenidos, ha devenido en un patrón de crecimiento exponencial que puede verse reflejado en la cantidad de resultados obtenidos tras realizar una búsqueda simple en cualquier motor de búsqueda.

Como respuesta al caos informativo, hace mucho que los bibliotecarios idearon estrategias de indización de contenidos temáticos y recuperación de la información, empleando vocabularios controlados que reflejaran el contenido de un libro sin lugar a ambigüedades y cuya estructura, compuesta por términos consensuados y relacionados entre sí, permitiera a los usuarios hallar documentos afines a su consulta inicial. Cabe señalar que, si bien este es el ideal, en la práctica los sistemas de información no suelen configurarse de forma tal que comprendan a la totalidad de relaciones entre los términos, lo cual puede traducirse en problemas como el ruido y el silencio informativo.

Mientras que esto ocurría en las bibliotecas, las peculiaridades de la virtualidad dieron lugar a otro tipo de escenarios. Fuera del ámbito bibliotecario y del alcance de profesionales de la información, los usuarios tuvieron que aplicar técnicas propias para describir los contenidos temáticos de los recursos que generaban, valiéndose de aquello que les resultaba intuitivo: el lenguaje natural.

En vista de las bondades de los vocabularios controlados, la simplicidad del medio que las personas emplean para comunicarse día a día puede parecer incompatible con este fin, especialmente considerando que la elección de descriptores puede responder a usos particulares del lenguaje o una visión muy personal del mundo. Sin embargo, este proceso rara vez ocurre de forma aislada, sino que se ve enriquecido gracias al contexto social en el que tiene lugar.

En medida que exista un grupo humano interactuando en un tiempo y espacio determinados, existirán también reglas tácitas que se han convertido en tales al ser aceptadas por la mayoría de la comunidad. Del mismo modo, un recurso etiquetado por varios usuarios contará con términos que predominarán sobre el resto gracias al número de veces que han sido empleados para describir su contenido, alcanzando cierto nivel de estabilidad semántica. De acuerdo con Kopeinik et al. (2017), este fenómeno se manifiesta en una creciente convergencia al escoger etiquetas para rangos particulares de temas. Mientras más estable sea un vocabulario, más útil será compartir recursos propios y aprovechar aquellos identificados por otros.

Para Wagner et al. (2014), una descripción estable semánticamente implica que los usuarios han llegado a un acuerdo implícito acerca de un conjunto de descriptores y su importancia en relación con un recurso, los cuales han de mantenerse estables a lo largo del tiempo. Como posibles causas de los procesos de estabilización semántica y creación de consenso en sistemas de etiquetado social, los autores mencionan la imitación del comportamiento de otros usuarios, los conocimientos previos compartidos y las propiedades intrínsecas del lenguaje natural. Es decir, esta combinación de factores de índole social lograría que una secuencia de etiquetas alcance un mayor grado de estabilidad de forma más rápida.

Adicionalmente, algunos servicios pueden recurrir a sistemas de recomendación que faciliten la identificación de los términos más frecuentes, así como a equipos de trabajo dedicados a la organización y desambiguación de estos, evocando la utilización de lenguajes documentales tradicionales.

Teniendo en cuenta que el etiquetado social puede llegar a presentar aspectos similares a la descripción de recursos de información en el ámbito bibliotecario, cabe preguntarse si los términos considerados como relevantes por los usuarios coinciden con los asignados por los catalogadores.

Definición del problema

En nuestro medio suelen emplearse catálogos en línea con características propias de la web 2.0 que, de encontrarse habilitadas, podrían permitir una mayor interacción entre los usuarios y la colección con que cuenta una unidad de información. Sin embargo, es usual que no se incluyan opciones adicionales como la posibilidad de asociar términos, así como herramientas que permitan a los usuarios agregar etiquetas a los recursos de información y elaborar listas personalizadas.

Si bien este panorama impide examinar el aporte de las etiquetas sociales a la recuperación de la información en un contexto más cercano, es posible recurrir a ejemplos provenientes de otras latitudes para establecer qué tan semejantes son los términos temáticos asignados por usuarios y catalogadores al momento de describir recursos de información. Ahora bien, explorar las posibles coincidencias entre el etiquetado social y la aplicación de un vocabulario controlado parte de establecer equivalencias entre ambos. En el caso de los encabezamientos de materia, las peculiaridades de su formato dificultan una comparación directa, por lo que es necesario aplicar técnicas propias de la minería de textos.

Contreras Barrera (2014) define la minería de textos como un proceso de descubrimiento de patrones en una colección de textos, para lo cual se sirve de la recuperación de información, el procesamiento de lenguaje natural, métodos estadísticos y matemáticos, entre otros. La construcción de dicha colección de textos, también denominada corpus, es el punto de partida para la extracción de información de interés.

Investigaciones previas suelen delimitar los corpus a examinar mediante la elección de una rama del conocimiento específica, a partir de la cual se seleccionan recursos de información que cuentan tanto con términos libres como controlados. De manera similar, la presente investigación pretende realizar una comparación entre las etiquetas de la comunidad virtual de catalogación LibraryThing y los encabezamientos de materia de la Library of Congress en inglés y español, centrándose en libros de ficción y no ficción publicados durante el periodo 1980-2019.

Con respecto a los encabezamientos de materia en inglés, se considerará a los términos seleccionados presentes en el registro bibliográfico de un recurso de información determinado, obviando a otros términos que, a pesar de guardar relaciones de jerarquía o equivalencia con los términos seleccionados en cuestión, no estén incluidos de manera explícita en el registro bibliográfico. En cuanto a los encabezamientos de materia en español, solamente se recurrirá al uso de un término general como reemplazo en caso de que no exista una equivalencia directa.

Objetivos

General

Comparar los términos asignados por los usuarios de LibraryThing y los catalogadores de la Library of Congress para la descripción temática de libros de ficción y no ficción.

Específicos

- Identificar patrones propios del lenguaje natural en la asignación de términos temáticos por parte de los usuarios de LibraryThing.
- Determinar la similitud entre los términos temáticos asignados por los usuarios de LibraryThing y los catalogadores de la Library of Congress, tanto en conjunto como por recurso de información.
- Determinar la asociación entre los términos temáticos asignados por los usuarios de LibraryThing y los catalogadores de la Library of Congress.

Justificación e importancia de la investigación

La producción académica alrededor del lenguaje natural se ha incrementado en los últimos años, desde interpretaciones de carácter lingüístico hasta casos de estudio relacionados con los negocios, pasando por análisis de sentimiento de las opiniones vertidas en redes sociales acerca de un tema específico. El etiquetado social tampoco ha sido ajeno a esta tendencia, y a pesar de que ha sido explorado desde diversas ópticas, es evidente que la mayoría de las investigaciones a nivel internacional provienen del ámbito de las ciencias de la computación y disciplinas afines.

Las tesis relacionadas con la minería de textos y el procesamiento de lenguaje natural a nivel nacional siguen la misma línea, incluso aquellas que buscan dar solución a problemas que atañen a la gestión de la información. En otras palabras, el enfoque suele darse a través de la informática en vez de la bibliotecología, hecho que puede explicarse gracias a la predominancia de los vocabularios controlados en nuestro campo, relegando el lenguaje natural a un rol menos significativo con respecto a la recuperación de la información.

En definitiva, la sola interrogante acerca de si el uso de etiquetas sociales es equiparable a los vocabularios controlados parece contravenir lo que nos resulta intuitivo en materia de análisis documental, especialmente teniendo en cuenta que la razón de ser de

estos últimos es evitar problemas como la sinonimia y la ambigüedad. Sin embargo, tiene sentido examinar las características del etiquetado social, no solamente desde el punto de vista de la descripción de contenidos, sino también como muestra del comportamiento informativo de los usuarios en contextos no tradicionales.

Desde el aspecto metodológico, se considera también pertinente ofrecer un ejemplo adicional del potencial de la ciencia de datos como herramienta de utilidad para la bibliotecología, sobre todo en relación con el análisis de datos no numéricos. Si bien la presente investigación solo se centra en ciertos aspectos de la minería de textos, las posibles aplicaciones de la ciencia de datos son múltiples y tienen como precedente a la integración de las tecnologías de la información y comunicación con el quehacer bibliotecológico, el cual responde ante todo a las necesidades de los usuarios.

Limitaciones de la investigación

En la búsqueda de una selección de recursos de información a partir de la cual se pudieran establecer ciertas generalizaciones, se realizó una exploración preliminar basada en la lista de ganadores del Premio Cervantes al considerarse como relevante dentro del ámbito de la cultura. Sin embargo, el número de libros descritos por la Library of Congress fue menor al esperado, por lo que se optó por usar como referencia a la lista de libros más vendidos del periódico The New York Times. Si bien esto garantizó que los libros seleccionados contaran tanto con etiquetas asignadas por los usuarios como con encabezamientos de materia asignados por los catalogadores, es de notar que el contenido refleja la realidad estadounidense.

Con respecto al idioma, mientras que el análisis de los términos en inglés pudo realizarse de forma directa, el análisis de los términos en español involucró el uso de las traducciones de las etiquetas sociales y los encabezamientos de materia a cargo de los usuarios de LibraryThing y la Dirección General de Bibliotecas de la Universidad Nacional

Autónoma de México, respectivamente. En el caso de las etiquetas sociales, su visualización en nuestro idioma es la opción por defecto al acceder a la versión en español de LibraryThing, aunque no siempre existen traducciones disponibles para todos los términos. En el caso de los encabezamientos de materia, la descripción realizada por la Library of Congress fue el punto de partida para la elaboración de una lista de términos en español, tomando como referencia al catálogo de autoridades LIBRUNAM. Si bien esto permitió comparar ambos conjuntos de datos, es importante notar que se trata de idiomas y contextos diferentes, lo cual impide establecer una equivalencia perfecta. En efecto, toda traducción involucra cierta pérdida de información, además de ser un procedimiento que difiere de la asignación directa de términos por parte de usuarios y catalogadores hispanohablantes.

Por último, el análisis no toma en cuenta equivalencias entre el significado de los términos, limitándose a encontrar semejanzas entre ambos conjuntos de datos en caso de que las grafías de las etiquetas sociales y de los encabezamientos de materia sean idénticas. Aunque se trata de una metodología validada por estudios similares, ello implica que algunos homógrafos podrían considerarse como términos coincidentes cuando no lo son en realidad, mientras que las etiquetas idénticas a formas no aceptadas de un encabezamiento no contarían como sinónimos.

Capítulo II: Revisión de literatura

Antecedentes del estudio

En los artículos *User-generated social tags versus librarian-generated subject headings: a comparative study in the domain of history* y *Measuring the applicability of user-generated social tags along with expert-generated LCSH descriptors in sociology: a heuristic study*, Samanta y Rath (2020, 2021) realizan una comparación entre las etiquetas de LibraryThing y los encabezamientos de materia de la Library of Congress en los campos de la historia y la sociología. Los autores indican que ambos vocabularios difieren debido a que los catalogadores buscan mejorar los puntos de acceso a las colecciones de la biblioteca, mientras que los usuarios emplean etiquetas para su propia recuperación de la información. Considerando la diferencia entre ambos, Samanta y Rath concluyen que las etiquetas sociales pueden mejorar la experiencia de los usuarios al complementar a los vocabularios controlados, mas no pueden reemplazar a estos últimos.

En el artículo *Supporting book search: a comprehensive comparison of tags vs. controlled vocabulary metadata*, Bogers y Petras (2017) emplean datos a gran escala de LibraryThing, Amazon, la Library of Congress y la British Library para el análisis. A pesar de que las etiquetas muestran mejores resultados, los vocabularios proveen de información complementaria durante la búsqueda. No obstante, ninguno de ellos sería completamente adecuado para responder a necesidades de información complejas.

En el artículo *Etiquetado social y blog-scraping como alternativa para la actualización de vocabularios controlados: aplicación práctica a un tesauro de biblioteconomía y documentación*, Mochón-Bezares et al. (2017) contrastan etiquetas tomadas de blogs especializados en ciencias de la información con descriptores e identificadores de la base de datos ISOC Biblioteconomía y Documentación. Los resultados muestran que las etiquetas cuentan con términos más variados y actualizados

que los listados de lenguaje controlado no estructurado, por lo que se sugiere que los primeros se tomen en cuenta al momento de incorporar nueva terminología al Tesauro de Biblioteconomía y Documentación del Centro de Información y Documentación Científica de España.

En el artículo *The comparative and analytical study of LibraryThing tags with Library of Congress Subject Headings*, Vaidya y Harinarayana (2016) señalan que, si bien hay ciertas coincidencias entre las etiquetas sociales y los encabezamientos de materia, las primeras muestran limitaciones inherentes al lenguaje no controlado. Aun así, se menciona que el resto de las etiquetas revelan una visión más amplia acerca de las obras, por lo que podrían mejorar la accesibilidad si se usaran en conjunción con los encabezamientos de materia.

En el artículo *Social tagging is no substitute for controlled indexing: a comparison of Medical Subject Headings and CiteULike tags assigned to 231,388 papers*, Lee y Schleyer (2012) comparan las etiquetas de CiteULike y los encabezamientos de temas médicos de la National Library of Medicine, aplicando diferentes técnicas de procesamiento de texto de forma progresiva. Concluyen que ambos grupos son marcadamente distintos en términos léxicos, aunque no necesariamente en términos semánticos, por lo que no consideran que las etiquetas puedan sustituir a los encabezamientos.

En la tesis *Social tagging versus expert created subject headings*, Rahman (2012) emplea un sistema de codificación propio para el análisis, encontrando que más de la mitad de las etiquetas de LibraryThing coincidían con los encabezamientos de materia de la Library of Congress, y que la frecuencia de uso de este grupo era mayor que la del resto de etiquetas. De acuerdo con el autor, ni las etiquetas ni los términos controlados funcionan de

forma satisfactoria por su cuenta, por lo que un catálogo híbrido sería capaz de ofrecer lo mejor de ambos mundos.

En el artículo *Exploring user-contributed metadata's potential to enhance access to literary works: social tagging in academic library catalogs*, DeZelar-Tiedman (2011) investiga si las etiquetas de LibraryThing asignadas a libros de literatura inglesa y estadounidense pueden servir de complemento a los encabezamientos de materia de la Library of Congress en el contexto de una biblioteca académica. Los resultados reflejan que si bien las obras clásicas y populares contaban con listas extensas de etiquetas, incluyendo términos específicos no incluidos dentro de los encabezamientos de materia, las etiquetas con mayor frecuencia de uso, así como aquellas asignadas a obras menos conocidas, tendían a ser más generales que los términos controlados, lo que no era de mucha utilidad para este tipo de bibliotecas.

En la ponencia *Contrasting controlled vocabulary and tagging: do experts choose the right names to label the wrong things?*, Heymann y Garcia-Molina (2009) se enfocan en la equivalencia sintáctica entre las etiquetas de LibraryThing y los encabezamientos de materia de la Library of Congress, encontrando que cerca de la mitad de los términos controlados coincidían con los no controlados. Sin embargo, la forma en que los usuarios y los catalogadores aplicaban dichos términos en común difería de manera considerable. Las explicaciones esbozadas por los autores —entre las que se encontraba el hecho de que los catalogadores no usan todos los términos que serían apropiados para la descripción de un recurso— sugerían que los encabezamientos de materia serían menos efectivos para la recuperación de los libros.

Marco teórico

Folksonomías

Tras un intercambio de ideas con Gene Smith y Eric Scheid —quienes se refirieron al etiquetado social como “un tipo de clasificación social informal" y "clasificación popular" (*folk classification*), respectivamente— la respuesta de Thomas Vander Wal fue: ¿Así que el desarrollo de una estructura categórica ascendente, creada por los usuarios y con un tesauro emergente se convertiría en una folksonomía? (Brandt & Medeiros, 2010; Morville & Rosenfeld, 2006).

Vander Wal (2007, como se citó en Peters, 2009) ahondó en el concepto al mencionar que lo que la gente realiza no es tanto categorizar sino establecer una forma de conectar elementos para otorgarles un significado de acuerdo con su propio entendimiento, y que dicho proceso tiene lugar en un entorno social. Por ende, se está haciendo referencia a dos niveles de actuación: uno es individual y está basado en la descripción de contenidos a través de términos que le son familiares a un usuario particular, mientras que el otro es colectivo e involucra el intercambio de información relacionado con recursos de información de interés común.

Siguiendo esa línea, Barros (2011) hace referencia a las folksonomías como un producto de la decisión de un usuario de organizar su propia información, utilizar su propio vocabulario y compartir su percepción de la clasificación con otros usuarios, una clasificación que puede ser elaborada y alterada en cualquier momento, y donde ocurre un proceso de retroalimentación inmediato al intercambiar las etiquetas menos usadas por otras que lo son más.

En efecto, los términos empleados por los usuarios pueden ir desde elecciones que guardan una relación obvia con el contenido hasta construcciones que evocan la tendencia aglutinante de los términos en *Newspeak*, el idioma ficticio de la obra *1984* de George Orwell. Sin embargo, en un sistema en el que la retroalimentación sea factible, la frecuencia de uso de términos no estandarizados será menor.

Dattolo et al. (2012) hacen una precisión terminológica adicional tomando como punto de partida a la definición de *etiqueta*. Esta vendría a ser un término escogido libremente por un usuario, considerado como significante para la descripción de un recurso, así como el mecanismo principal para navegar y buscar nuevos recursos en sistemas de etiquetado social. La colección de todas las etiquetas asignadas por un usuario sería equivalente a su personomía, mientras que la colección de todas las personomías presentes en un sistema sería denominada folksonomía.

Gómez-Díaz (2013) incide en esta distinción al mencionar que:

> Una folksonomía es simplemente el conjunto de términos donde personas que usan un mismo código (vocabulario) esperan encontrar de nuevo el mismo objeto. Gracias a las etiquetas se crea un espacio constituido por las aportaciones de todos los usuarios de determinados servicios, sin una intervención centralizada ni más autoridad que el uso que hagan de estas los propios usuarios (Sección de Los resultados, párr. 6).

En conclusión, si bien las folksonomías tienen como origen a una práctica individual, en realidad son la suma de múltiples casos enmarcados en un contexto que permite el intercambio de ideas entre los miembros de una comunidad. Basta con que cada usuario conceda su aprobación a las propuestas de representación de la información formuladas por otros usuarios o establezca mecanismos de acceso adicionales que sean capaces de integrarse a los preexistentes.

Sistemas de etiquetado social

Para Marinho et al. (2012), los sistemas de etiquetado social son aplicaciones de la web 2.0 enfocadas en la publicación y etiquetado de recursos web por usuarios de internet ordinarios, tareas a través de las cuales emerge cierto sistema de clasificación colaborativo

denominado como folksonomía. Sin embargo, la arquitectura y las funcionalidades de los sistemas de etiquetado social varían de sistema a sistema.

Según Zubiaga et al. (2011), estas diferencias dan lugar al etiquetado colaborativo y al etiquetado simple, lo cual resulta en folksonomías amplias y estrechas, respectivamente. En el primero de los casos, múltiples usuarios etiquetan un mismo recurso con sus propias etiquetas y a través de su propio vocabulario, siendo perfectamente posible que más de un usuario emplee un mismo término para la descripción. Mientras tanto, en el segundo caso, uno o pocos usuarios asignan etiquetas únicas. Como se desprende de lo dicho por Klašnja-Milićević et al. (2017), a diferencia de las folksonomías amplias, un mismo término no puede estar asociado a un mismo recurso múltiples veces.

Delicious fue el primero en introducir un sistema de marcadores basado en etiquetas (Halpin, 2013), siendo uno de los ejemplos más notables de folksonomías amplias durante los años en que estuvo en operación. Al momento de registrar un nuevo marcador, Delicious recomendaba etiquetas basadas en la descripción realizada por otros usuarios, además de permitir la visualización y descubrimiento de otros recursos asociados a una etiqueta específica, un tipo de navegación conocido como *pilot browsing* (Anderson, 2012).

Por su parte, Flickr, usualmente mencionado en relación con la creación de folksonomías estrechas, permite el etiquetado social pero con ciertas restricciones como la no duplicidad de términos (Kipp et al., 2017). Al ser únicas, las etiquetas se encuentran asociadas directamente con un recurso, lo que las hace particularmente relevantes para la recuperación de objetos que podrían ser difíciles de encontrar o que solo pueden ser recuperados si una descripción textual está asociada con el recurso (De Meo et al., 2013), como es el caso de las fotografías.

Análisis temático del contenido

Guimarães et al. (2007) hacen referencia al análisis documental de contenido como un conjunto de procedimientos de naturaleza analítico-sintética, los cuales comprenden el análisis del contenido temático de los documentos y su síntesis, sirviéndose de lenguajes documentales para su representación y posterior recuperación por parte de los usuarios. Para los autores, dicha representación consiste en expresar el contenido temático de un documento de forma estandarizada y de acuerdo a parámetros previamente establecidos, lo que la asemeja a la traducción.

Tal como mencionan Chowdhury et al. (2008), al tratarse de un lenguaje precoordinado, la representación del contenido a través del uso de encabezamientos de materia implica el uso de una lista de términos organizados de acuerdo a un orden predeterminado. En el caso específico de los encabezamientos de materia de la Library of Congress, Steele (2012) señala además ciertos principios a ser considerados por los catalogadores, tales como la asignación de uno o más encabezamientos que sinteticen de mejor manera el contenido del recurso y brinden acceso a sus temas más importantes; la especificidad, entendida como un concepto que refleja la relación particular entre el encabezamiento y el recurso; y la objetividad, que implica considerar la intención del autor o editor sin emitir juicios de valor.

En cuanto a la representación del contenido a través de etiquetas sociales, el proceso es similar en esencia, aunque difiere en algunos aspectos. Stock y Stock (2013) señalan que las folksonomías son una especie de sistema de etiquetado libre, donde los usuarios de los documentos asignan palabras claves sin considerar ningún tipo de regla. Ante la problemática generada por la falta de lineamientos a seguir, Peters (2009) propone recurrir a métodos de desambiguación y estructuración retroactivas a cargo de los mismos usuarios. Según la autora, estas actividades no se opondrían al diseño liberal y orientado al

usuario de las folksonomías, sino que responderían a las necesidades personales de los usuarios de un sistema.

Ventajas de las folksonomías sobre los vocabularios controlados

De acuerdo con Klašnja-Milićević et al. (2017), aunque el etiquetado realizado por expertos —como es el caso de los catalogadores— asegura la calidad de los resultados, este implica un proceso largo y costoso. Por otro lado, el etiquetado automatizado no requiere de la intervención humana, pero puede producir etiquetas con un bajo nivel de precisión, además de requerir un uso intensivo de datos. En contraposición a ambos, el etiquetado social tiene la ventaja de producir colecciones de etiquetas a gran escala cuya calidad tiende a aumentar en relación con el número de usuarios activos de un sistema.

Precisamente, Stock y Stock (2013) mencionan el bajo costo como una de las ventajas de las folksonomías, lo cual facilita el manejo de grandes bases de datos que difícilmente podrían haber sido catalogadas de otra forma, así como el uso auténtico del lenguaje por parte de los usuarios, incluyendo la pronta adopción de neologismos.

Teniendo en cuenta la ingente cantidad de recursos generados día a día, sería imposible asignar términos controlados desde un órgano centralizado, por lo que la delegación de esta función a los usuarios cobra sentido. Asimismo, considerando que la actualización de un tesauro requiere de un consenso previo, la incorporación de nuevos términos tampoco podría darse con la misma rapidez.

Además del bajo costo de las folksonomías, Batley (2014) considera la serendipia —hallazgo imprevisto, pero de gran relevancia— como una de sus fortalezas, específicamente en relación con la navegación a través de las nubes de etiquetas. Aunque este tipo de visualización facilita el descubrimiento de recursos de forma espontánea, al carecer de la estructura provista por los esquemas de clasificación y las taxonomías, las nubes de etiquetas no muestran ninguna relación conceptualmente significativa entre los

términos. Para la autora, sin embargo, el sentido de comunidad y participación sería el rasgo más importante del etiquetado colaborativo: mientras que la categorización en las taxonomías formales es impuesta desde arriba y suele reflejar las prioridades de la organización a cargo, las folksonomías responden directamente a las necesidades de los usuarios.

Gilton (2016) destaca la inclusividad de las folksonomías, la cual es posible gracias a que las barreras de acceso a los sistemas de etiquetado social suelen ser mínimas. Los usuarios pueden crear cuentas gratuitas y generar sus propios metadatos con facilidad, especialmente considerando que la curva de aprendizaje de las folksonomías es mucho menor a la de las taxonomías. Como consecuencia, las folksonomías reflejan la forma en que la gente común encuentra, usa y categoriza la información. Los términos de búsqueda son relevantes porque son los mismos usuarios quienes los suministran, lo cual no impide que sean de utilidad al público en general.

Adicionalmente, el etiquetado cumple con otros roles que no necesariamente se encuentran divorciados de su razón de ser, sino que son una expresión más de su naturaleza social. Es así que las etiquetas pueden desempeñar una gran variedad de funciones comunicativas más allá de su capacidad para organizar o categorizar temas (Wikström, 2014).

Desventajas de las folksonomías frente a los vocabularios controlados

Para Stock y Stock (2013), las desventajas de las folksonomías se deben principalmente a su falta de precisión. Al no contar con mecanismos de control terminológico, estas suelen contener erratas, juicios de valor, términos sincategoremáticos —conceptos incompletos que solo cobran significado en relación con otros términos— y descripciones de acciones planeadas. Asimismo, los autores sostienen que los usuarios no siempre distinguen entre la descripción del contenido y la descripción formal, mezclando

distintos niveles semánticos como el *ofness*, el *aboutness* y la iconología, especialmente en el caso de documentos no textuales. Es decir, se trata de la diferencia entre etiquetar un libro electrónico con el término "ciencia" para aludir a su temática y asignar el término "libro electrónico", lo que podría ser innecesario a pesar de representar la naturaleza del objeto.

Hjørland (2017) sostiene que *aboutness* debe considerarse como sinónimo de "materia" en el ámbito de las ciencias de la información, mientras que *ofness* suele emplearse para referirse a objetos o eventos que tengan lugar en una imagen. Por su parte, la *iconología* implica un trasfondo cultural que ayuda a distinguir, por ejemplo, a una estatua de la Venus de Milo. Aunque no es tan común como los otros conceptos, *isness* tiene que ver más con la forma que con el fondo.

Zeng et al. (2011) dan el ejemplo de una pintura de un atardecer (*ofness*) en San Francisco (*aboutness*) para referirse a la representación genérica y específica del contenido de una obra, respectivamente. *Isness*, por otro lado, trata de la forma o el género, tal como novela, sonata, mapa o fotografía. Son estos conceptos los que los usuarios aplican indistintamente y de manera altamente subjetiva al momento de describir contenidos, algunos de los cuales pueden ser poco relevantes para la recuperación de la información dependiendo del contexto.

De acuerdo con Gupta et al. (2011), la ambigüedad, otro de los problemas relacionados con las folksonomías, se presenta debido a que diferentes usuarios asignan términos de distinta manera, empleando desde acrónimos hasta combinaciones de palabras en una sola etiqueta y sin mediar espacios. Incluso en el caso de que una etiqueta esté compuesta por una sola palabra, la falta de contexto puede llevar a una interpretación incorrecta de su significado. Los autores advierten además la presencia de sinónimos,

homónimos y formas singulares y plurales de una misma palabra, lo cual puede explicarse por la falta de lineamientos formales para el etiquetado.

Gilton (2016) menciona debilidades como la falta de jerarquías y relaciones entre términos, las dificultades para lidiar con palabras compuestas, frases, plurales, variantes ortográficas, palabras polisémicas y términos imprecisos o poco relevantes. Asimismo, usuarios malintencionados pueden recurrir al *tag spam*, asignando etiquetas que inducen al error a otros usuarios acerca del contenido de los recursos con el objetivo de que visiten sitios inapropiados. En vista de ello, la autora concluye que los puntos débiles de las folksonomías se dan en áreas donde las taxonomías tienen éxito y viceversa. Por tanto, las folksonomías no reemplazarían ni a la descripción bibliográfica ni a los vocabularios controlados, sino que serían un complemento que permitiría maximizar las posibilidades de acceder a recursos de información.

Etiquetado social en bibliotecas

Porter (2011) se refiere a las folksonomías como agentes de descubrimiento y recuperación dinámicos y centrados en el usuario que, sin embargo, carecen de la precisión necesaria para reemplazar a las taxonomías formales. Diversos estudios coinciden en este punto, considerando que el etiquetado puede asistir en la búsqueda de información, pero es preferible contar con un mayor nivel de control terminológico al momento de la descripción. De todas maneras, se estima que su integración con los catálogos de biblioteca es beneficiosa para los usuarios, ya que pueden disponer tanto de la versatilidad de las folksonomías como de la fiabilidad de los vocabularios controlados.

De acuerdo con Fernández Ramos (2017), los principales motivos por los que las folksonomías tienen un lugar en los catálogos son, por un lado, el cierre de la brecha entre el vocabulario que emplean los usuarios y el vocabulario de los lenguajes documentales y, por otro lado, la posibilidad de que los usuarios organicen la información según sus propias

necesidades. Asimismo, las etiquetas son útiles para describir ciertos recursos de forma más específica, tal como es el caso de las obras de ficción y documentos cuyos temas aún no han sido incorporados a los encabezamientos de materia.

Ahora bien, habilitar el etiquetado en los catálogos de biblioteca no siempre garantiza que los usuarios harán uso de esa funcionalidad. Johansson y Golub (2019) sostienen que, considerando que un número relativamente grande de etiquetas es necesario para que surja un consenso con respecto a la descripción de un recurso, puede ser conveniente para las bibliotecas acudir a una fuente externa como LibraryThing for Libraries. Este servicio potencia los catálogos a través de la importación de etiquetas que han pasado por un proceso de control previo, aunque no siempre cuentan con la uniformidad esperada.

Fuera de los catálogos, las bibliotecas también han interactuado con los sistemas de etiquetado social de forma directa para sacar partido de la colaboración masiva (*crowdsourcing*). Kipp et al. (2017) mencionan el caso de Flickr, utilizado por instituciones culturales alrededor del mundo para que los usuarios interesados puedan enriquecer la información asociada a sus colecciones fotográficas mediante la asignación de etiquetas y anotaciones adicionales.

By the People, un proyecto de la Library of Congress, y Comunidad BNE de la Biblioteca Nacional de España funcionan de manera similar, invitando a los usuarios a transcribir y etiquetar documentos históricos con el objetivo de incrementar su usabilidad (Sánchez Nogales, 2020; Zwaard et al., 2018). Como comentan Zwaard et al., este tipo de intercambios permiten que las bibliotecas entablen un diálogo con la comunidad, facilitando el descubrimiento y fomentando el uso a partir de una experiencia más personal con sus colecciones.

Estrategias de mejora del etiquetado social

A lo largo del tiempo se han planteado múltiples alternativas que buscan subsanar las deficiencias presentadas por las folksonomías, desde buenas prácticas hasta procesos que involucran a seres humanos y máquinas en distinta medida. Siguiendo esa línea, Stock y Stock (2013) hacen referencia al *tag gardening*, término acuñado por Isabella Peters y Katrin Weller para referirse al procesamiento de las etiquetas sociales para optimizar la recuperación de la información. Las estrategias originadas a partir de este concepto comprenden la edición y eliminación de etiquetas, la implementación de sistemas de recomendación, el control del vocabulario, la interacción con otros sistemas de organización del conocimiento y la delimitación de las etiquetas usadas con mayor frecuencia, también denominadas como *power tags*.

Sistemas de recomendación. Hoy en día los sistemas de recomendación son ubicuos, pudiendo observarse su presencia en redes sociales, plataformas de video bajo demanda, mercados de comercio electrónico, así como en sistemas de etiquetado social. Según Belém et al. (2017), a diferencia de los sistemas de recomendación generales que buscan hacer sugerencias de acuerdo con los intereses de los usuarios, la recomendación de etiquetas busca además describir, resumir y organizar el contenido de un ítem. Los autores mencionan seis métodos de recomendación representativos: aquellos basados en la coocurrencia de etiquetas, en el contenido de los recursos (usados comúnmente para mitigar la ausencia de etiquetas iniciales), en la factorización de matrices, en grafos, en técnicas de agrupamiento y en el aprendizaje de clasificación.

Entre ellos, FolkRank es uno de los métodos basados en grafos con mayor rendimiento. Desarrollado a partir del algoritmo PageRank de Google, FolkRank está basado en un grafo no dirigido con enlaces representando relaciones coocurrentes entre nodos de usuarios, contenidos y etiquetas. A pesar del nivel de precisión de sus resultados, no está carente de problemas, tales como el alto costo computacional al generar

recomendaciones (Gedikli, 2013; Wang et al., 2019). Por otro lado, aunque utiliza las relaciones entre usuarios, ítems y etiquetas de manera efectiva, no considera las relaciones internas de usuario-usuario, ítem-ítem y etiqueta-etiqueta, lo cual le impide hacer uso del comportamiento informacional de usuarios extremadamente similares al usuario objetivo (Zhao et al., 2021).

Otro de los fenómenos relacionados con los sistemas de recomendación es el arranque en frío (*cold start*), un problema que surge cuando el sistema no puede hacer deducciones acerca de usuarios o ítems sobre los cuales aún no cuenta con información suficiente (Zaccone & Karim, 2018). Los autores mencionan que una solución frecuente es optar por un enfoque híbrido entre el emparejamiento basado en el contenido y el filtrado colaborativo. De esta manera, los ítems que aún no han sido clasificados por los usuarios obtendrían una calificación automática basada en las calificaciones asignadas por la comunidad a otros ítems con contenidos similares.

Formalización de relaciones entre etiquetas. Si bien el etiquetado social no cuenta con estructuras jerárquicas semejantes a las de los vocabularios controlados, sí es posible observar ciertos tipos de relaciones entre los términos a partir del análisis de un conjunto de etiquetas. Shafique et al. (2019) mencionan las relaciones de subsunción, similitud, coocurrencia y equivalencia, conceptos que tienen su correspondencia en el ámbito lingüístico.

Según Shafique et al. (2019), una relación de *subsunción* implica que una etiqueta abarca a otra conceptualmente. Siempre que se etiquete a un recurso con el término “aritmética” se podrá usar también “matemática”, pero lo opuesto no es necesariamente cierto. *Similitud* implica que ambas etiquetas poseen características comunes a nivel semántico, mientras que *coocurrencia* alude a etiquetas que aparecen juntas con frecuencia. *Equivalencia* hace referencia a dos etiquetas que comparten el mismo

significado, tales como una palabra en singular y su forma en plural o un sustantivo y su abreviatura correspondiente.

Este tipo de relaciones no son inmediatamente aparentes, por lo que existen métodos que se basan en la heurística, los recursos léxicos externos y el aprendizaje automático para extraer etiquetas relevantes y organizarlas en formas de conocimiento estructurado (Dong et al., 2018). Para los autores, los métodos heurísticos utilizan reglas para definir e inferir relaciones, pero no definen formalmente las relaciones semánticas entre etiquetas ni funcionan de forma adecuada si los datos son escasos. Los métodos basados en recursos léxicos externos recurren a estos para encontrar relaciones entre términos, pero se ven perjudicados por su cobertura limitada. Por su parte, los métodos basados en el aprendizaje automático usan técnicas supervisadas o no supervisadas para descubrir patrones jerárquicos, pero puede que no discriminen entre términos subordinados, relacionados y paralelos.

Evidentemente, a pesar de que los algoritmos utilizados para recrear estructuras formales continúan siendo perfeccionados, aún tienen ciertas limitaciones. Ante ello, algunos servicios han optado por recurrir a sistemas híbridos que incluyen el control manual de las folksonomías generadas por los usuarios.

Curaduría de folksonomías. Bullard (2019) emplea este concepto para referirse al proceso que utiliza el conjunto de etiquetas creadas por los usuarios como insumo y, a través de la toma de decisiones expertas o colectivas, identifica y mitiga problemas de sinonimia y homografía. De acuerdo con la autora, los usuarios de la comunidad virtual de catalogación de libros LibraryThing pueden, por ejemplo, identificar qué etiquetas pueden considerarse como variantes y cuáles no, dando lugar a versiones normalizadas de los términos. Stack Overflow, un sitio de preguntas y respuestas sobre programación, cuenta además con páginas de información similares a las notas de alcance de los vocabularios

controlados, las cuales son editadas por usuarios previamente designados como buenos contribuidores. Por su parte, el repositorio de fanficción Archive of Our Own dispone de voluntarios que normalizan las etiquetas empleadas en las obras derivadas y agrupan algunos términos bajo otros más generales, creando así relaciones jerárquicas.

Para Aghaebrahimian et al. (2020), las categorías de la enciclopedia en línea Wikipedia son el resultado de la negociación entre los editores, y si bien la taxonomía resultante es generalmente intuitiva, posee deficiencias que disminuyen su utilidad para el etiquetado de colecciones de textos. Ante ello, los autores proponen soluciones relacionadas con el ecosistema de Wikipedia, tales como el uso de Wikidata —una base de conocimientos editada colaborativamente— como vocabulario controlado. Como menciona Feliciati (2022), a diferencia de Wikipedia, Wikidata preserva la información como datos estructurados dentro de una base de datos. Asimismo, cuenta entre sus usuarios a grupos de bibliotecarios y archivistas que colaboran de forma activa con el enriquecimiento de los metadatos.

Definición de términos

- Bolsa de palabras: Modelo de representación del texto de un documento a través de palabras sueltas, ignorando la posición de las palabras y la estructura de las oraciones en favor de la frecuencia con que cada palabra aparece en el texto (Bernico, 2018; Müller & Guido, 2017). Denominado en inglés como *bag of words*.
- Cadena de caracteres: Tipo de datos que permite almacenar una secuencia de caracteres como letras, números y signos de puntuación en una sola variable (Meza Vega, 2018). Denominado en inglés como *string*.
- Conjunto de datos: Serie de elementos de datos aproximadamente equivalentes a una hoja de cálculo bidimensional o una tabla de una base de datos (Tchakounté & Hayata, 2017). Denominado en inglés como *dataset*.

- Etiquetas idiosincráticas: Etiquetas consideradas como significativas y útiles por un solo usuario (Wu et al., 2006, como se citó en Derham & Mills, 2010).
- Encabezamientos de materia: Puntos de acceso a un asiento bibliográfico, compuestos de una palabra o frase que designa el contenido temático de un recurso de información. Se encuentran enlazados a otros encabezamientos de materia a través de referencias cruzadas, expresadas a través de términos generales y específicos (Policy and Standards Division, 2013; H. Young, 1988).
- Expresiones regulares: Secuencia de caracteres que permiten construir patrones empleados para realizar tareas de sustitución de caracteres, realizar búsquedas de patrones en una cadena dada y construir funciones de validación (Caballero González, 2016). Denominadas en inglés como *regular expressions* o *regex*.
- Lenguaje de programación: Conjunto de reglas, palabras, símbolos y códigos usados para escribir programas de computadora (Morley & Parker, 2012).
- Literatura de ficción: Categoría de obras literarias en las que prevalecen elementos narrativos inventados (Iglesias Rebollo & González Gordon, 2005).
- Literatura de no ficción: Categoría de obras literarias que versan sobre hechos reales en contraposición a obras de carácter ficticio (Iglesias Rebollo & González Gordon, 2005).
- Palabras vacías: Palabras de alta frecuencia, tales como pronombres, determinantes y preposiciones (Ignatow & Mihalcea, 2017). Denominadas en inglés como *stop words*.
- Raspado web: Recolección de datos que por lo general se da a través de un programa automatizado, el cual se encarga de realizar consultas a un servidor web y solicitar datos en forma de código HTML y otros archivos de los que está compuesta una página web, además de analizar los datos obtenidos para extraer la

información que se necesita (Mitchell, 2015). Denominado en inglés como *web scraping*.

- Selector CSS: Dentro de una regla definida en una hoja de estilo en cascada, parte que especifica qué elementos HTML se verán afectados por la regla. Como ejemplo, el selector `a` hace referencia al elemento `<a>`, equivalente a los enlaces de una página web (Dorman, 2020).
- Similitud léxica: Semejanza basada en la sintaxis, estructura y contenido de un conjunto de textos. A diferencia de la similitud semántica, no considera el significado de las palabras ni su contexto (Sarkar, 2019).
- Término específico: Encabezamiento de materia que representa la especie, la parte o un ejemplo de un concepto más amplio. Designa a un miembro de la clase representada por un término general, con el que guarda una relación de tipo jerárquico (Martínez Tamayo & Mendes, 2015; Policy and Standards Division, 2013).
- Término general: Encabezamiento de materia que representa el género o el todo. Designa la clase a la que pertenece un término específico, con el que guarda una relación de tipo jerárquico (Martínez Tamayo & Mendes, 2015; Policy and Standards Division, 2013).
- Término no seleccionado: Sinónimo o cuasisinónimo de un término seleccionado, con el que guarda una relación de equivalencia. Denominado también como término no preferente, no puede asignarse en la indización, pero sirve para reenviar la consulta hacia un término preferente (Martínez Tamayo & Mendes, 2015; Velasco de Diego, Llorens Morillo & Moreiro González, 1999).
- Término relacional: Asociación conceptual entre encabezamientos de materia. Designa una relación entre dos términos vinculados de una forma distinta a la

jerárquica (Martínez Tamayo & Mendes, 2015; Policy and Standards Division, 2013).

- Término seleccionado: Término designado para representar un concepto general y que sirve para describir el contenido de un documento y realizar consultas. Denominado también como término preferente al haber sido seleccionado sobre otros sinónimos que pudieran corresponder (Martínez Tamayo & Mendes, 2015; Velasco de Diego, Llorens Morillo & Moreiro González, 1999).
- Tokenización: Proceso por el cual se separa un texto determinado en frases, palabras, símbolos u otros elementos significativos, denominados como tokens (Kumar & Paul, 2016).
- XML (*Extensible Markup Language*): Estándar abierto usado para codificar y describir datos estructurados, así como para facilitar el mantenimiento, organización, intercambio y reutilización de los datos por programas de computadora (Cole & Han, 2013).

Estrategias y técnicas de investigación

La técnica empleada para la elaboración del marco teórico y la posterior selección de los materiales y métodos fue el análisis documental. Asimismo, y como parte de ese proceso, la revisión de investigaciones previas llevó a la elección de LibraryThing como la comunidad virtual de catalogación por analizar, especialmente considerando su predominancia en los estudios dedicados a comparar etiquetas sociales y encabezamientos de materia, así como la posibilidad de contar con equivalentes en nuestro idioma de las etiquetas asignadas a cada título mediante la consulta de la versión en español de su sitio web.

Búsqueda y recuperación de la información

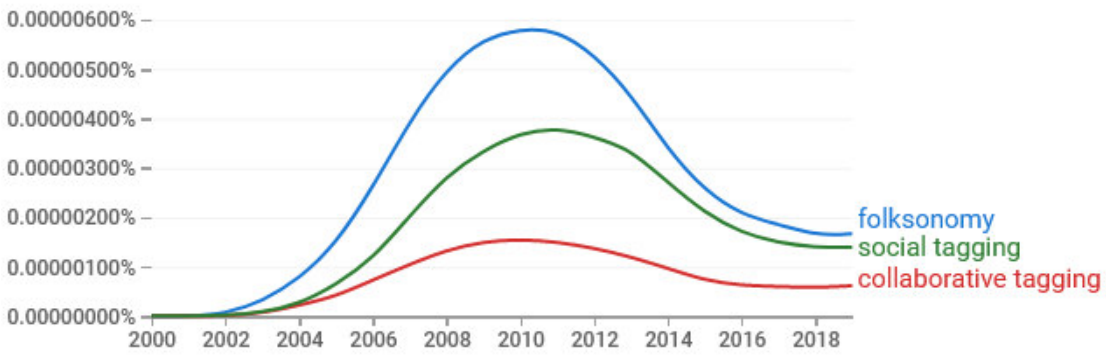
Además de Google Books, se empleó a Google Scholar como punto de partida para la búsqueda de información, accediendo a repositorios especializados como E-LIS, aquellos asociados a universidades como DiVA, y aquellos asociados a editoriales académicas como Taylor & Francis Online.

Las palabras clave incluyeron términos en inglés como *collaborative tagging* (etiquetado colaborativo), *folksonomy* (folksonomía), *social tagging* (etiquetado social), *social tagging systems* (sistemas de etiquetado social), *subject headings* (encabezamientos de materia), *tags* (etiquetas) y *text mining* (minería de textos), empleándose las comillas para realizar una búsqueda exacta y combinando términos en caso de que fuera necesario un nivel mayor de precisión.

Salvo contadas excepciones, se siguieron los lineamientos respecto a la antigüedad de las fuentes de información. En el caso de la presente investigación, dicha delimitación temporal coincide parcialmente con la relevancia del objeto de estudio en materia de publicaciones. Efectivamente, y tal como muestra la Figura 1, el corpus de libros impresos de Google Ngram Viewer (Michel et al., 2011) indica que los términos *folksonomy*, *social tagging* y *collaborative tagging* cobraron mayor relevancia entre los años 2010 y 2011.

Figura 1

Folksonomía, etiquetado social y etiquetado colaborativo en Google Ngram Viewer



Criterios de elección de la información

El uso de Google Scholar facilitó la identificación de artículos científicos, permitiendo además la visualización del número de citas recibidas. Si bien no fue un criterio de exclusión, en caso de encontrar artículos científicos similares se dio prioridad al que contara con un mayor índice de citas. Adicionalmente, se consideró de manera preferente a:

- Documentos en inglés y español.
- Documentos publicados durante el periodo 2011-2021.
- Documentos provenientes de repositorios asociados a instituciones o editoriales académicas.
- Documentos a texto completo, considerándose también documentos con visualización parcial en caso de que la sección en cuestión sea de libre acceso.
- Documentos cuyo título, mención de responsabilidad y datos de publicación estén expresamente señalados.

Se excluyeron documentos cuyo enfoque no estaba relacionado con la organización de la información sino que trataban sobre otros fenómenos de carácter social. A pesar de que se encuentran fuera del alcance de la presente investigación, es oportuno mencionar dentro de los temas de interés asociados con las folksonomías al procesamiento de lenguaje natural en la recuperación de información mediante el uso de inteligencia artificial, el desarrollo de comunidades virtuales gracias a la interacción entre los usuarios de un servicio, y la función comunicativa de las etiquetas en el marco de movimientos sociales, facilitando la planificación y el intercambio de información alrededor de las movilizaciones y dando lugar a la emergencia de una narrativa alrededor de los hechos.

Capítulo III: Hipótesis y variables

Hipótesis

General

Los términos asignados por los usuarios de LibraryThing muestran coincidencias parciales con los asignados por los catalogadores de la Library of Congress, pero son lo suficientemente distintos para considerarlos como un complemento en la descripción temática.

Específicas

- Los términos temáticos asignados por los usuarios de LibraryThing presentan palabras vacías, términos ambiguos y términos no relacionados con el contenido de forma directa.
- El grado de similitud entre los términos temáticos asignados por los usuarios de LibraryThing y los catalogadores de la Library of Congress es bajo.
- El grado de asociación entre los términos temáticos asignados por los usuarios de LibraryThing y los catalogadores de la Library of Congress es moderado.

Variable de estudio

Similitud entre los términos asignados por los usuarios de LibraryThing y los catalogadores de la Library of Congress.

Operacionalización de la variable de estudio

Tabla 1

Matriz de operacionalización

Variable	Dimensiones	Indicadores	Escala [a, b, c]	
			Valor	Interpretación
Similitud	Similitud léxica	Coeficiente de similitud	0	Similitud nula
			1	Similitud perfecta
		Radio de cobertura	0	Cobertura nula
			1	Cobertura perfecta
	Asociación entre términos	Coeficiente de correlación	0	Correlación nula
			±0,2 - ±0,4	Correlación débil
			±0,6 - ±0,9	Correlación fuerte
			±1	Correlación perfecta

Nota. [a] Torres-Moreno (2014, p. 298). [b] Lee y Schleyer (2012, p. 1752). [c] Ñaupas Paitán et al. (2014, p. 263).

Capítulo IV: Materiales y métodos

Área de estudio

LibraryThing

LibraryThing (https://www.librarything.com) es una comunidad virtual de catalogación de libros que facilita la interacción entre los usuarios mediante grupos de discusión, reseñas, recomendaciones y listas. Su funcionalidad principal, sin embargo, es la elaboración de colecciones personales que permiten a los usuarios llevar un inventario de los libros que están leyendo, que han leído o que planean leer.

Cada obra cuenta con géneros denominados GenreThing (ver Tabla 2), los cuales son asignados automáticamente por LibraryThing basándose en diversas fuentes, incluyendo etiquetas y términos controlados usados por bibliotecas y librerías. Los usuarios pueden realizar cambios a los géneros asignados según sea necesario, los cuales pueden ser adoptados para su visualización por parte de la comunidad en general tras un análisis previo.

La organización de los libros puede darse también a través de etiquetas que consisten en palabras o frases definidas por los usuarios separadas por comas al momento de su ingreso, por lo que admiten el uso de espacios en blanco y caracteres especiales (ver Figura 2).

A diferencia de los géneros, los cuales pueden agrupar diversas etiquetas, las

Tabla 2

Géneros asignados por LibraryThing para la obra Cien años de soledad

Géneros	Traducción al español
Fantasy	Fantasía
Fiction and literature	Ficción y literatura

etiquetas carecen de relaciones jerárquicas. No obstante, cada etiqueta suele contar con etiquetas relacionadas que han sido marcadas como sinónimos por los usuarios, así como con traducciones en diversos idiomas. Si bien las traducciones tuvieron como punto de partida a Wikipedia, los usuarios también participan del proceso añadiendo nuevas traducciones y validando las traducciones existentes a través de un sistema de votos (ver Tabla 3).

Figura 2

Nube de etiquetas de LibraryThing para la obra Cien años de soledad

Tabla 3

Etiquetas más usadas en LibraryThing para la obra Cien años de soledad

Etiquetas	Traducción al español provista por LibraryThing
Fiction	Ficción
Magical realism	Realismo mágico
To-read	Por leer
Novel	Novela
Literature	Literatura
Latin America	Latinoamérica
Colombia	Colombia
Latin American literature	Literatura latinoamericana
Spanish	Español
South America	América del Sur

Library of Congress

Reconocida como la biblioteca nacional de los Estados Unidos de América, la Library of Congress (https://loc.gov) usa como vocabularios controlados a los encabezamientos de materia (campos 650 y 651) y género/forma (campo 655) (ver Figura 3). Aunque las subdivisiones de forma continúan en uso, los encabezamientos de género/forma son producto de la decisión de la Library of Congress de separarlos de los encabezamientos de materia con el objetivo de describir qué es una obra en vez de describir de qué trata (J. Young, 2009), un ejemplo clásico de *isness* y *aboutness*.

Adicionalmente, y aunque no son considerados como puntos de acceso, los registros bibliográficos pueden incluir otros vocabularios controlados, tales como BISAC, una lista de descriptores empleada por editoriales norteamericanas, FAST, un proyecto de colaboración entre el OCLC y la Library of Congress, y MeSH de la National Library of Medicine, así como listas de encabezamientos de materia y tesauros provenientes de otros países (Network Development and MARC Standards Office, 2021).

A pesar de que los encabezamientos de materia de la Library of Congress no

Figura 3

Encabezamientos de la Library of Congress para la obra Cien años de soledad, vista de etiquetas MARC

650 _0 |**a** Macondo (Imaginary place) |**v** Fiction.
651 _0 |**a** Latin America |**x** Social conditions |**v** Fiction.
655 _7 |**a** Epic literature. |**2** gsafd

Nota. El código *gsafd* indica que la fuente del encabezamiento de género/forma son las *Guidelines on subject access to individual works of fiction, drama, etc.* de la American Library Association.

cuentan con una traducción oficial al español, su adaptación para la descripción de los materiales documentales de la Universidad Nacional Autónoma de México —a cargo de la Dirección General de Bibliotecas (https://www.dgb.unam.mx)— convierte al catálogo de autoridades LIBRUNAM en una fuente idónea para la búsqueda de términos equivalentes en español.

Si bien LIBRUNAM emplea fuentes adicionales para la elaboración de sus registros de autoridades, cada término aceptado incluye entre sus variantes la versión original en inglés del encabezamiento proveniente de la Library of Congress. Incluso en el caso de que un término no cuente con un registro individual, también es posible encontrarlo dentro de la lista de términos específicos asociados a un término más general o como una materia asignada de forma directa a un registro bibliográfico (ver Tabla 4).

Durante el análisis, y con el objetivo de establecer equivalencias que permitan determinar la similitud entre los términos, tanto las etiquetas sociales como los encabezamientos se considerarán como términos aislados, evocando la usanza de los lenguajes poscoordinados.

Diseño de investigación

De acuerdo con Argimon Pallás y Jiménez Villa (2019), las características más importantes de la arquitectura de un estudio pueden clasificarse según ejes, como la finalidad del estudio, la secuencia temporal y el control de la asignación de los factores de estudio.

En cuanto a su finalidad, la presente investigación tiene un diseño descriptivo al no tener como objetivo el análisis de una relación causal, sino que se limita a utilizar los datos disponibles de forma puramente descriptiva. Asimismo, su secuencia temporal es transversal debido a que los datos corresponden a un momento específico en el tiempo,

Tabla 4

Términos asociados con el tema "realismo mágico" según el catálogo de autoridades LIBRUNAM

Tipos de términos	Términos
Término aceptado	Realismo mágico (Literatura)
Término no aceptado	*Magic realism (Literature)*
Término relacionado	Magia en la literatura
	Lo maravilloso en la literatura
	Realismo en la literatura
Término general	Novela fantástica
	Surrealismo
Término en inglés	*Magic realism (Literature)*

mientras que su clasificación en términos del control de la variable de estudio, por parte de la investigadora, la convierte en observacional.

Población y muestra

La población está conformada por los libros más vendidos durante el periodo 1980-2019, en conformidad con The New York Times. Siguiendo la distinción por formato, tipo de literatura y público que hace el periódico, las listas que se usaron de referencia están restringidas a libros de tapa dura de ficción y no ficción dirigidos al público adulto (Hawes Publications, 2021).

A continuación, se realizó una búsqueda en el catálogo de la Library of Congress, a partir de la cual se excluyó a los libros cuyos registros no contaban con encabezamientos de materia. Como resultado se obtuvo un total de 620 libros de ficción y 430 libros de no ficción. Se optó por mantener a ambos grupos por separado para comprobar si la temática incide en la similitud entre las etiquetas sociales y los encabezamientos de materia.

Con el objetivo de determinar una muestra apropiada para ambos grupos, se recurrió a la fórmula $n = \frac{z^2 p(1-p)}{e^2}$, donde *n* es el tamaño de la muestra, *z* es el valor Z correspondiente al nivel de confianza, *p* es la proporción para la cual se calcula el margen de error y *e* es el margen de error. Por convención, *p* suele ser igual a 0,5, lo que equivale a un resultado conservador. Si se considera a 0,5 como valor por defecto, la fórmula se convierte en $n = \frac{z^2}{4e^2}$. Para calcular una proporción con un nivel de confianza de 95% y un margen de error de 5%, se obtiene (Laffly, 2020):

$$n = \frac{z^2}{4e^2} = \frac{(1{,}96)^2}{4(0{,}05)^2} = 384{,}16 \approx 385$$

Por último, la selección de los 385 libros de ficción y los 385 libros de no ficción correspondientes se dio de forma aleatoria, la cual sirvió tanto para el análisis de los términos en inglés como de los términos en español.

Procedimientos, técnicas e instrumentos de recolección de datos

Las búsquedas preliminares en LibraryThing y en el catálogo de la Library of Congress se realizaron manualmente. Para el resto de los procedimientos se empleó a R (R Core Team, 2021), un lenguaje de programación y entorno para computación estadística y gráficos, y RStudio, un entorno de desarrollo integrado para R.

Al ser un proyecto de código abierto, R cuenta con una gran cantidad de paquetes adicionales dedicados a propósitos específicos (Hornik, 2020). Dichos paquetes potencian las funciones básicas del software y permiten la recolección, el análisis y la visualización de datos.

Recolección de datos

Términos en inglés. La recolección de datos tuvo lugar a través del raspado web, una técnica para extraer datos de un sitio web de forma automatizada. Primeramente, se especificó de qué páginas se iba a extraer la información, para lo cual se partió de la

recopilación manual del identificador asignado por LibraryThing para cada título y del número de control asignado por la Library of Congress para cada registro bibliográfico.

A diferencia de los identificadores, la cantidad de números de control fue mayor al número de títulos debido a que se consideraron todos los registros bibliográficos que contaban con asientos secundarios de materia. Como resultado se obtuvo un total de 545 registros para libros de ficción y 579 registros para libros de no ficción.

Para la manipulación de datos se empleó a dyplr, un paquete de R que permite añadir, seleccionar, sintetizar, ordenar y agrupar variables, y que se usó durante las diferentes etapas del presente análisis. Tanto dyplr como rvest y xml2 forman parte de tidyverse, una colección de paquetes que engloba las tareas propias de todo proyecto de ciencia de datos, tales como la importación, limpieza, manipulación y visualización de datos, así como elementos de programación (Wickham et al., 2019).

En el caso de LibraryThing, rvest fue necesario para capturar el código HTML de las páginas seleccionadas, así como para especificar qué selector CSS correspondía a las etiquetas (Wright et al., 2021). Asimismo, debido a que las etiquetas cargaban junto con la página a través de JavaScript, se tuvo que recurrir a PhantomJS, un navegador sin interfaz gráfica que simula la interacción con una página web por parte de un usuario.

Si bien LibraryThing permite la visualización de todas las etiquetas asociadas a un título, se decidió limitar la recuperación a las etiquetas que aparecen en la visualización por defecto, lo cual corresponde a las 30 etiquetas más usadas, salvo en los casos en que un título cuente con un número menor de etiquetas asignadas.

En el caso de la Library of Congress, no fue posible extraer la información de la misma manera debido a que no había un selector CSS de uso exclusivo para los encabezamientos de materia. Por tanto, se recurrió a xml2 para recuperar los registros en

MARCXML, un esquema que permite aplicar la estructura de XML a todo tipo de registros MARC 21 (Network Development and MARC Standards Office, 2020).

Como nota adicional, durante el proceso de raspado web se añadió una función de espera al código a efectos de no sobrecargar a los servidores con múltiples solicitudes por minuto. Para ello se inspeccionó el archivo robots.txt para tomar conocimiento de los lineamientos y restricciones establecidos por cada sitio web.

Términos en español. Teniendo en cuenta que los identificadores para cada título no varían en la versión en español de LibraryThing, el procedimiento fue similar para el caso de las etiquetas en español. Solamente bastó con modificar el dominio de nivel superior de la dirección web durante el nuevo proceso de recolección de datos, intercambiando ".com" por ".es". En cuanto a los encabezamientos de materia, se recurrió al preprocesamiento de datos para delimitar la selección a partir de la cual se realizaría la búsqueda manual de términos equivalentes en el catálogo de autoridades LIBRUNAM.

Preprocesamiento de datos

Además de dyplr, esta etapa requirió del uso de stringr, un paquete de tidyverse dedicado a la manipulación de cadenas de caracteres.

Términos en inglés. En el caso de LibraryThing, el preprocesamiento involucró la eliminación de etiquetas usadas por un solo usuario y aquellas que indicaban el precio de los libros. Asimismo, se reemplazó el término *non-fiction* por *nonfiction* con el objetivo de evitar la pérdida de información durante el proceso de tokenización.

En el caso de la Library of Congress, se determinó que los campos MARC a extraer de los registros bibliográficos serían los de nombre personal (campo 600), nombre corporativo (campo 610), nombre de la reunión (campo 611), título uniforme (campo 630), término temático (campo 650), nombre geográfico (campo 651) y término de género/forma (campo 655). La inclusión de campos adicionales al de término temático se debió a que los

registros bibliográficos no siempre contaban con el campo 650, sino que se limitaban a señalar un nombre personal, corporativo o geográfico, acompañado, en algunos casos, de una subdivisión como es el caso de *fiction*.

Debido a que la estructura de MARCXML difiere de la de XML, la recuperación del contenido de los campos no se realizó de forma directa sino mediante el uso de expresiones regulares que permitieron extraer la información dentro de las etiquetas MARCXML.

Una vez obtenidos los campos, se restringió la selección de los encabezamientos de materia a aquellos cuyo segundo indicador fuera 0 (Lista de encabezamientos de la Library of Congress) o 1 (Lista de encabezamientos de la Library of Congress para literatura infantil), excluyendo a los términos provenientes de otras fuentes. Adicionalmente, se eliminaron espacios extra y caracteres especiales, así como los encabezamientos duplicados por título.

Términos en español. Durante la etapa de preprocesamiento se detectaron algunas diferencias con las etiquetas en inglés. Esto debido a que la recolección de las etiquetas en español se realizó en una oportunidad distinta, lo que explica ciertas variaciones con respecto a las etiquetas más usadas y su frecuencia de uso. Adicionalmente, las etiquetas que no contaban con equivalentes en español se mantuvieron en su idioma original.

El preprocesamiento en sí fue similar al de las etiquetas en inglés, por lo que se empezó excluyendo a las etiquetas idiosincráticas y aquellas que indicaban el precio de los libros. Adicionalmente, se corrigieron errores de codificación que no permitían visualizar los signos diacríticos de forma adecuada y se reemplazó a los caracteres “o/a” por “o” en etiquetas como “Clásico/a”. Del mismo modo, se reemplazó al término “no ficción” por “noficción” al considerar que también era un término relevante para el análisis en nuestro idioma.

En el caso de las autoridades de nombre, la mayoría de los términos se dejaron tal cual, salvo en los casos en que existiera una traducción provista por el catálogo de autoridades LIBRUNAM. De no existir un equivalente directo, el criterio para reemplazar un término en inglés por su equivalente en español fue considerar a un término más general, mientras que en el caso de los nombres geográficos se incluyeron topónimos cuya traducción era conocida. Solo unos cuantos términos se mantuvieron en inglés al no contar con una traducción oficial.

Tokenización

Wickham (2014) define a los datos ordenados (*tidy data*) de acuerdo con tres principios basados en la relación entre el contenido de un conjunto de datos y su estructura: cada variable corresponde a una columna, cada observación corresponde a una fila, y cada tipo de unidad de observación corresponde a una tabla. Siguiendo esa línea, Silge y Robinson (2021) se refieren a los datos ordenados aplicados al análisis de texto como una tabla con un token por fila.

Para esta etapa se requirió de tidytext (Silge & Robinson, 2016), un paquete para la minería de textos basado en los principios de *tidy data*. Por defecto, la tokenización mediante el uso de tidytext divide el texto en palabras sueltas, elimina signos de puntuación y convierte los tokens en minúsculas.

También se recurrió a stopwords (Benoit et al., 2021), un paquete que contiene listas de palabras vacías para diferentes idiomas provenientes de diversas fuentes. En el caso de la presente investigación, los glosarios empleados fueron snowball y nltk, tanto para los términos en inglés como para los términos en español.

Términos en inglés. Tanto en el caso de LibraryThing como en el de la Library of Congress, se separaron los términos en tokens y se excluyeron a aquellos que coincidían con la lista de palabras vacías. Posteriormente se eliminaron signos de puntuación

restantes, tokens que terminaban en un número y tokens con menos de tres caracteres al considerarse como poco significativos, así como palabras vacías adicionales basándose en los tokens más frecuentes por cada conjunto de datos. Por último, se eliminaron los tokens duplicados por título.

En cuanto a los números, su eliminación se debió a que no se esperaban coincidencias entre ambos conjuntos de datos. Por un lado, los números en LibraryThing usualmente hacen referencia al año de publicación o al año en que un usuario leyó un libro en particular, mientras que la Library of Congress suele emplear números para indicar fechas asociadas a una autoridad de nombre personal o para delimitar periodos históricos de manera precisa. No obstante, sí se conservó a números ordinales que hicieran referencia a siglos, ya que en este caso el uso sí coincidiría entre ambos conjuntos de datos.

Como siguiente paso, Lee y Schleyer (2012) recurrieron a un algoritmo de *stemming* empleado por varios sistemas de recuperación de texto y motores de búsqueda para reducir el nivel de variación entre palabras. Empero, se consideró que, si bien la técnica es de gran utilidad en ese ámbito, era más relevante para la presente investigación el contar con la forma completa de las palabras para realizar una comparación más precisa.

Términos en español. La tokenización fue similar para el caso de las etiquetas en español. Luego de separar los términos en tokens y excluir a las palabras vacías, se eliminaron signos de puntuación restantes, tokens que terminaban en un número y tokens con menos de tres caracteres, excepto por el número en romanos “XX”, de modo que pudiera contextualizarse el uso de la palabra “siglo” de ser necesario. Antes de eliminar los tokens duplicados por título, se eliminaron palabras vacías adicionales basándose en los tokens más frecuentes por cada conjunto de datos, los cuales coincidieron en su gran mayoría con las palabras vacías adicionales en inglés.

Análisis estadístico

Los estudios realizados sobre el tema emplean diferentes técnicas para establecer el grado de similitud y correlación entre los términos (ver Tabla 5). Entre los métodos para comparar la similitud entre textos, Li et al. (2020) mencionan aquellos basados en palabras clave como el coeficiente de Jaccard y aquellos basados en modelos de espacio vectorial como la similitud de coseno. Mientras que Jaccard puede calcularse a partir de un modelo de bolsa de palabras (Srinivasa-Desikan, 2018), la similitud de coseno requiere convertir el texto en un vector en el espacio mediante técnicas como TF-IDF, que mide qué tan importante es una palabra para un documento en el marco de una colección de documentos (Li et al., 2020; Silge & Robinson, 2021). Adicionalmente, el coeficiente de Jaccard no se ve afectado por términos duplicados, lo que sí ocurre con la similitud de coseno (Campesato, 2021).

Algunos autores también utilizan el radio de cobertura como medida de similitud

Tabla 5

Medidas de similitud aplicadas en el análisis de textos

Autores	Medidas de similitud
Heymann y Garcia-Molina (2009)	Similitud de Jaccard
	Correlación de rango de Kendall
Zubiaga et al. (2011)	Similitud de coseno
Lee y Schleyer (2012)	Similitud de Jaccard
	Radio de cobertura
Rahman (2012)	Correlación de Pearson
Vaidya y Harinarayana (2016)	Similitud de Jaccard
	Radio de cobertura
Samanta y Rath (2020, 2021)	Similitud de Jaccard
	Correlación de rango de Spearman
Silge y Robinson (2021)	Correlación de Pearson

(Lee & Schleyer, 2012; Vaidya & Harinarayana, 2016). Dados un conjunto de etiquetas y un conjunto de términos controlados, los radios de cobertura de las etiquetas y de los términos controlados serán equivalentes a las anotaciones en común divididas entre el total de etiquetas y el total de términos controlados, respectivamente. Para Lee y Schleyer (2012), el radio de cobertura puede ayudar a determinar si los encabezamientos pueden ser sustituidos por las etiquetas o viceversa. Por ejemplo, los encabezamientos podrían convertirse en sugerencias durante el etiquetado en caso de que abarquen una proporción significativa de las etiquetas. Por otro lado, las etiquetas podrían ser un aporte valioso a la descripción en caso de que no aparezcan de forma frecuente dentro de los encabezamientos.

En cuanto a las medidas de correlación, el coeficiente de Pearson calcula qué tanto varían dos variables de manera conjunta y luego contrasta el resultado con qué tanto varían por sí mismas. Esta medida se basa en ciertos supuestos como el principio de homocedasticidad, según el cual los puntos se encuentran distribuidos de forma uniforme a lo largo de una línea recta. Si la distribución incluye valores atípicos que no se ajustan a la línea, entonces el coeficiente de Pearson no es el más apropiado para medir la asociación entre dos variables. Ante ello debe recurrirse a los coeficientes de Spearman o de Kendall, los cuales se basan en los rangos de los valores. De entre ellos, es mejor emplear el coeficiente de Kendall ante un gran número de empates entre los rangos para evitar distorsiones en los resultados (Hinton et al., 2014).

Coeficiente de similitud de Jaccard

Para el cálculo se empleó a textreuse (Mullen, 2020), un paquete de herramientas para medir la similitud entre documentos y detectar qué pasajes fueron reutilizados. El coeficiente de similitud de Jaccard es el resultado de la intersección de dos conjuntos dividida por su unión:

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

Radio de cobertura

El radio de cobertura se obtuvo a través de radio_of_matches, otra de las funciones de textreuse (Mullen, 2020). El resultado de la función es el número de elementos en *B* que también están en *A*, dividido por el número total de elementos en *B*:

$$Radio\ de\ cobertura(B) = \frac{A \cap B}{B}$$

Si bien ratio_of_matches computa los elementos duplicados, al trabajar con términos únicos se logra el mismo resultado que los ejemplos de radio de cobertura dados por Lee y Schleyer (2012).

Coeficiente de correlación de rango de Kendall

Para el cálculo se empleó a cor, una de las funciones básicas de R que permite el cómputo de los coeficientes de Pearson, Spearman y Kendall. De acuerdo con Sarabia Alegría y Pascual Sáez (2012), considerando un total de *n* parejas de elementos para ordenar, así como un número de concordancias *C* y un número de discordancias *D*, el coeficiente τ de Kendall se obtiene de la siguiente forma:

$$\tau = \frac{C - D}{\frac{n(n-1)}{2}}$$

Visualización de datos

Las figuras fueron creadas en su gran mayoría a través de ggplot2 (Wickham, 2020), un paquete que toma como insumo un conjunto de datos y construye representaciones visuales de los datos a través de objetos geométricos, escalas, paneles y sistemas de coordenadas. También se empleó a wordcloud (Fellows, 2018) para visualizar nubes de palabras y a scales (Wickham & Seidel, 2020) para controlar aspectos relacionados con la presentación de los diagramas de dispersión.

Capítulo V: Resultados

Presentación y análisis de los resultados

El análisis comprende cuatro conjuntos de datos por idioma (inglés y español), los cuales pueden clasificarse según su naturaleza (etiquetas sociales y encabezamientos de materia) y el tipo de literatura de que tratan (ficción y no ficción). Se consideró pertinente analizar escenarios similares y aplicar los mismos procedimientos con el fin de determinar si los resultados guardaban semejanza independientemente del idioma de los términos, así como para identificar las cualidades propias de ambos grupos.

Primero se detallan las características generales de los conjuntos de datos luego del raspado web, así como las modificaciones que tuvieron lugar a partir del preprocesamiento de datos. A continuación, y de acuerdo con los objetivos trazados, se procede a identificar patrones en la asignación de etiquetas, incluyendo instancias señaladas por diversos autores como deficiencias asociadas con las folksonomías.

Asimismo, se precisa cuáles son los tokens de uso más frecuente dentro del grupo de etiquetas y encabezamientos relacionados con un tipo de literatura determinado, así como su frecuencia en el conjunto de datos opuesto. Posteriormente, se analizan los valores del coeficiente de similitud de Jaccard y el radio de cobertura para cada conjunto de datos, al igual que por cada uno de los títulos.

Por último, se observan las coincidencias entre las frecuencias de uso de los términos de acuerdo con el algoritmo y posterior análisis propuesto por Silge y Robinson (2021), así como la interpretación de los valores del coeficiente de correlación de rango de Kendall.

Etiquetas en inglés

El conjunto de datos de ficción estuvo conformado por 11 550 etiquetas, de las cuales 1729 resultaron ser únicas. Mientras tanto, el conjunto de datos de no ficción estuvo

conformado por 11 536 etiquetas, de las cuales 3145 resultaron ser únicas. La mayoría de títulos contaron con las 30 etiquetas que aparecían en la visualización por defecto, salvo en el caso de dos títulos de no ficción con 20 y 26 etiquetas, respectivamente.

Como puede observarse en las nubes de etiquetas (ver Figuras 4 y 5), los conjuntos de datos incluyeron ejemplos de falta de uniformidad en el uso de mayúsculas, términos ambiguos como abreviaturas y siglas, términos en otros idiomas como *roman* (novela, en alemán), sinónimos como *current affairs* y *current events* (temas de actualidad), descripciones de acciones planeadas como *read* (leído) y *to-read* (por leer), referencias acerca del formato como *ebook* (libro electrónico) y *paperback* (encuadernación en rústica), así como referencias acerca del ejemplar como *first edition* (primera edición) y *signed* (firmado).

Figura 4

Etiquetas en inglés más frecuentes, ficción

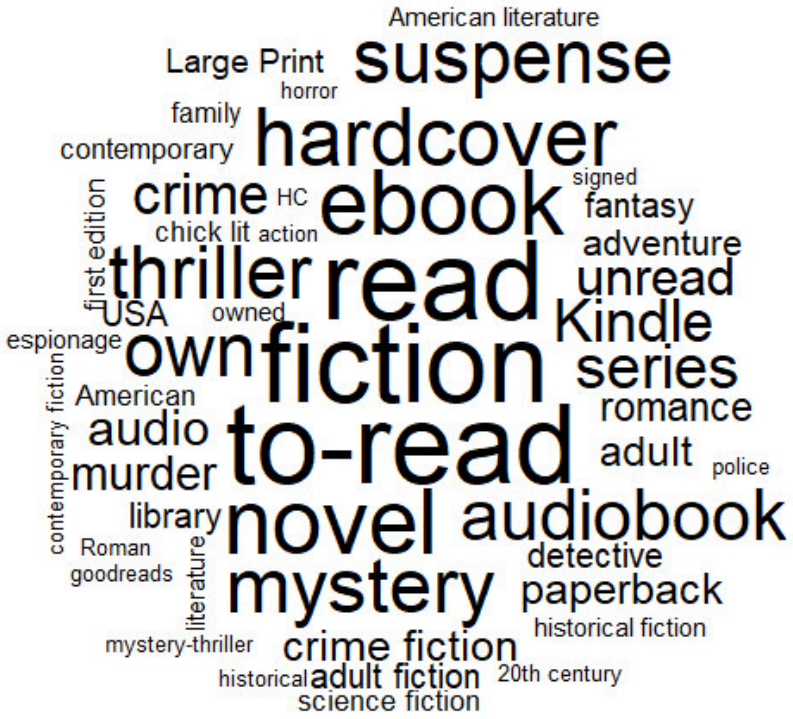
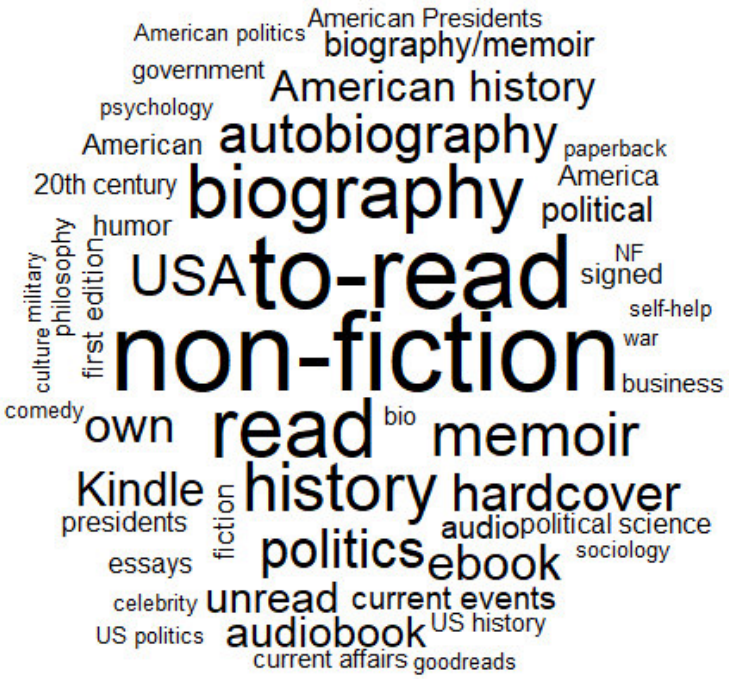


Figura 5

Etiquetas en inglés más frecuentes, no ficción



En contraposición a los problemas observados, también pudieron advertirse términos que hacían referencia al contenido de forma específica. Es así que diversos géneros literarios destacaron en el caso de los títulos de ficción, mientras que, en el caso de los títulos de no ficción, la biografía como género apareció de forma prominente, así como términos que aludían al espíritu de la época.

Como parte del preprocesamiento de datos, la eliminación de las etiquetas usadas por un solo usuario y aquellas que indicaban el precio de los ejemplares tuvo efectos distintos en ambos conjuntos de datos. En el caso de los títulos de ficción, el número total de etiquetas se redujo en 1,1%, mientras que las etiquetas únicas se vieron reducidas en 5,4%. En el caso de los títulos de no ficción, la disminución fue más notoria debido a que el número total de etiquetas se redujo en 9,4%, mientras que las etiquetas únicas se vieron

reducidas en 27,3%. Ello puede explicarse debido a que los títulos de no ficción presentaron un grado mayor de diversidad en términos de contenido, por lo que la presencia de etiquetas idiosincráticas también fue mayor.

Encabezamientos en inglés

Tras la recolección de datos y la selección de los campos 600, 610, 611, 630, 650, 651 y 655 —correspondientes al formato MARC 21— cuya procedencia fuera la Lista de encabezamientos de la Library of Congress, el conjunto de datos de ficción estuvo conformado por 1493 encabezamientos complejos, de los cuales 810 resultaron ser únicos. Mientras tanto, el conjunto de datos de no ficción estuvo conformado por 1509 encabezamientos complejos, de los cuales 1112 resultaron ser únicos. Al igual que en el caso de los conjuntos de datos de LibraryThing, los encabezamientos asignados a los títulos de no ficción contaron con un grado mayor de variabilidad.

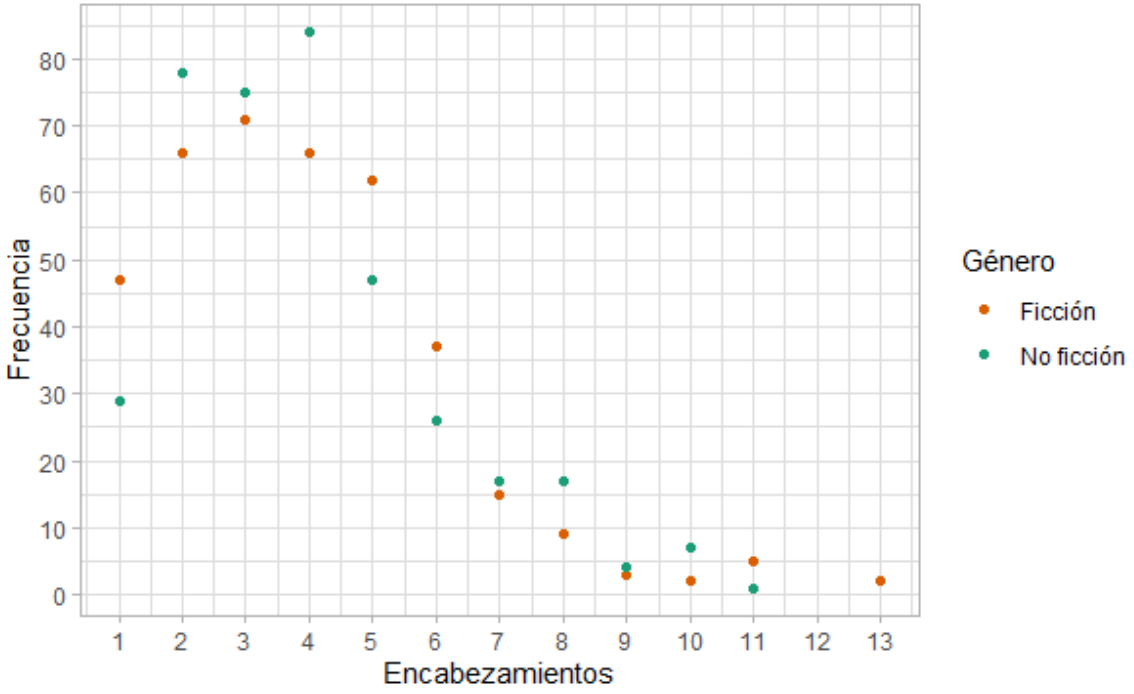
En cuanto al número de encabezamientos por registro, la mayoría de ellos contaban con dos a cuatro encabezamientos (ver Figura 6). Las frecuencias fueron similares tanto en el caso de los títulos de ficción como en los de no ficción, lo que indica que no hubo grandes diferencias entre ambos conjuntos de datos en cuanto a la descripción. En todo caso, la distinción se dio entre el número de las etiquetas y los encabezamientos, siendo las primeras más numerosas que los últimos, como era previsible.

Tokens en inglés

Tras la tokenización, y continuando con el preprocesamiento de datos, la eliminación de las palabras vacías adicionales a las del paquete stopwords se dio en función a las etiquetas. Esto debido a que, a diferencia de los encabezamientos, las etiquetas contaban con términos no relacionados con el contenido dentro de los tokens más frecuentes. Los términos considerados como no significativos incluyeron descripciones de acciones planeadas y referencias acerca del formato, principalmente.

Figura 6

Encabezamientos en inglés por registro



Tal como puede observarse en la Tabla 6, el porcentaje en que se redujo el número de tokens fue mayor en aquellos provenientes de las etiquetas de ficción, mientras que lo opuesto ocurrió con los tokens provenientes de los encabezamientos de no ficción. Sin embargo, como puede observarse en la Tabla 7, esto no tuvo un efecto considerable en el número de tokens únicos.

Como resultado del preprocesamiento, el promedio de tokens por título fue de 20,1 y 21,7 en el caso de las etiquetas de ficción y no ficción, respectivamente, mientras que en el caso de los encabezamientos de ficción y no ficción, el promedio de tokens por título fue de 9,1 y 10,0, respectivamente.

Frecuencia. Con respecto a los tokens más frecuentes en los conjuntos de datos de ficción, aquellos provenientes de los encabezamientos mostraron un énfasis en el género, correspondiente a la subdivisión de forma *fiction* y los diversos términos de género/forma

Tabla 6

Tokens en inglés antes y después del preprocesamiento

Tipo de literatura	Antes	Después	Reducción
Etiquetas			
Ficción	15442	7728	50,0%
No ficción	14119	8359	40,8%
Encabezamientos			
Ficción	5245	3506	33,2%
No ficción	6675	3837	42,5%

Tabla 7

Tokens únicos en inglés antes y después del preprocesamiento

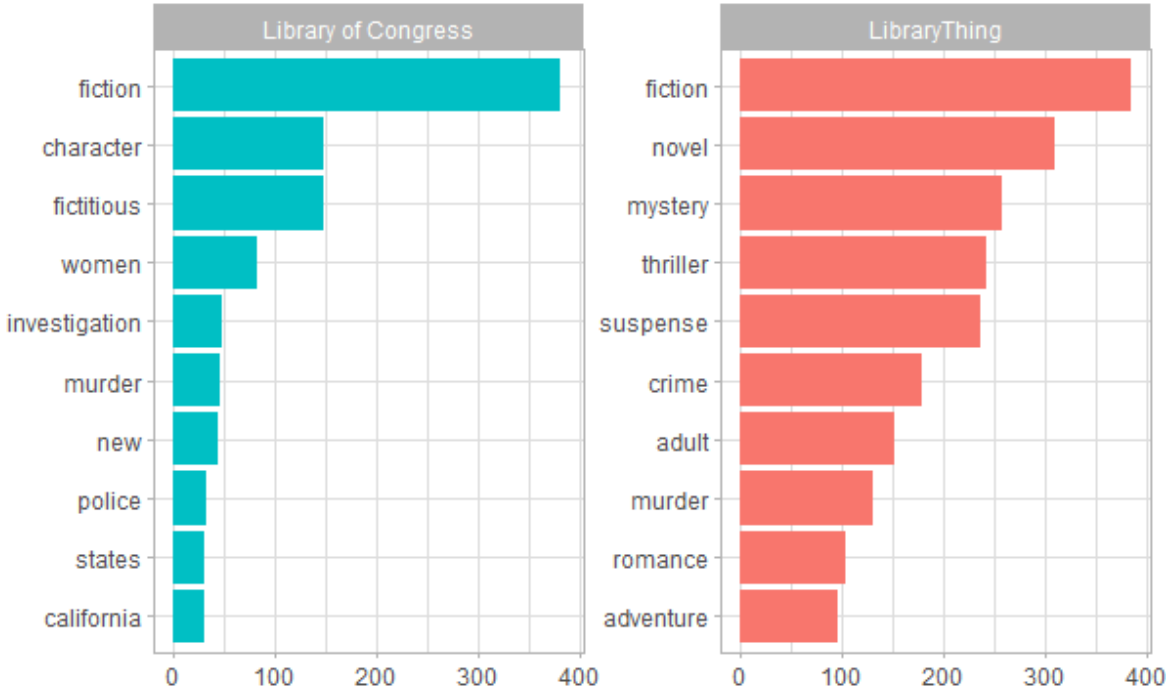
Tipo de literatura	Antes	Después	Reducción
Etiquetas			
Ficción	1556	1365	12,3%
No ficción	2107	1917	9,0%
Encabezamientos			
Ficción	1001	922	7,9%
No ficción	1397	1223	12,5%

que incluyen esta palabra. Asimismo, el hecho de que los términos que componen el calificativo *fictitious character* (personaje ficticio) estuvieran entre los tokens más frecuentes refleja la práctica de asignar como materias a los personajes de un título dado (ver Figura 7).

En general, en ambos conjuntos destacó la prevalencia que se le da a la descripción de los géneros literarios, destacando aquellos de corte policial. Asimismo, los tokens provenientes de los encabezamientos incluyeron algunos topónimos que originalmente se emplearon junto con la subdivisión de forma *fiction*, los cuales correspondían a nombres

Figura 7

Tokens en inglés más frecuentes, ficción



corporativos (campo 610), nombres geográficos (campo 651) y subdivisiones geográficas (subcampo $z), principalmente.

Por otro lado, y como puede observarse en la Tabla 8, los términos usados en los encabezamientos se encontraban también en las etiquetas, lo cual no siempre ocurrió de forma opuesta. Nuevamente, esto puede explicarse debido a que las etiquetas eran mucho más numerosas, por lo que abarcaron un mayor número de términos no contemplados en los encabezamientos de materia. Destacó, sin embargo, el grado de coincidencia entre ambos conjuntos de datos con respecto a la frecuencia de uso del término *fiction*.

Con respecto a los conjuntos de datos de no ficción, los términos más usados en los encabezamientos correspondieron al topónimo *United States*, cuyo equivalente en el conjunto de datos opuesto fue la sigla USA. La frecuencia de uso evidenció que, si bien la temática se centró alrededor de los Estados Unidos de América en ambos casos, las

Tabla 8

Frecuencia de los tokens en inglés más frecuentes, ficción

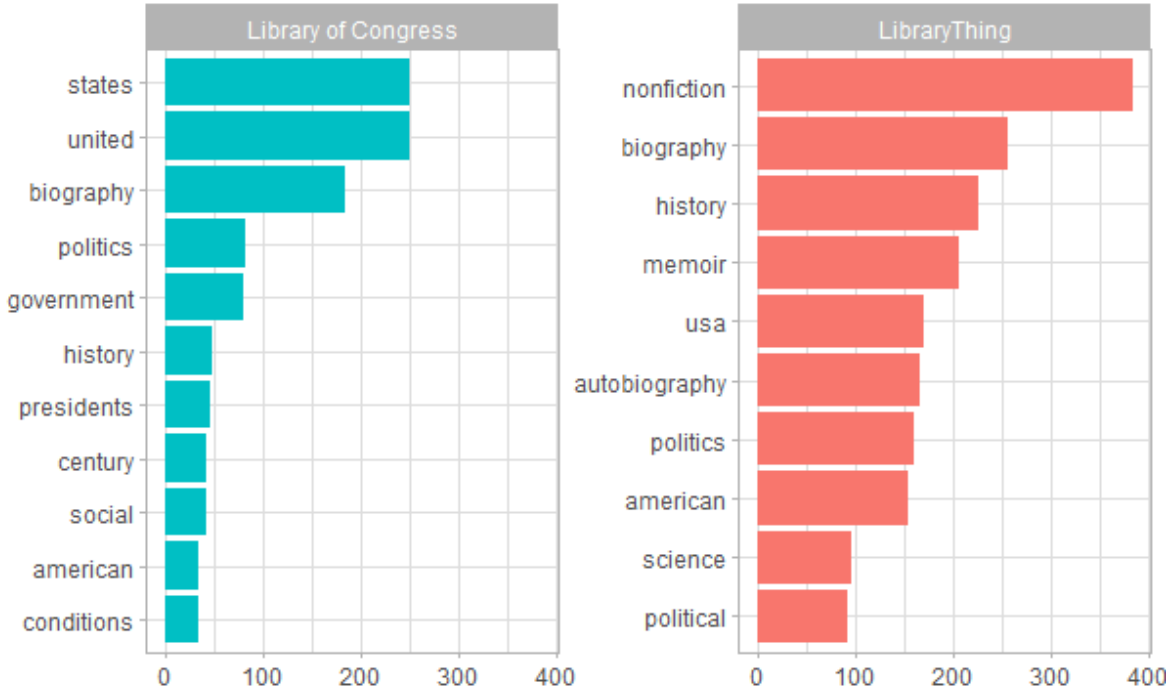
Encabezamientos			Etiquetas		
Tokens	Frecuencia	Frecuencia en etiquetas	Tokens	Frecuencia	Frecuencia en encabezamientos
fiction	380	385	*fiction*	385	380
character	147	6	*novel*	310	-
fictitious	147	6	*mystery*	257	5
women	82	16	*thriller*	243	-
investigation	47	3	*suspense*	236	3
murder	46	131	*crime*	178	2
new	43	58	*adult*	151	4
police	32	49	*murder*	131	46
states	31	1	*romance*	103	-
california	30	23	*adventure*	96	1

etiquetas dieron prioridad a *nonfiction*, término sin equivalente directo dentro de los tokens más frecuentes provenientes de los encabezamientos. Sin embargo, ambos conjuntos de datos coincidieron en el uso de los términos *american* (estadounidense) *biography* (biografía), *history* (historia) y *politics* (política). Asimismo, el término *science* (ciencia) destacó como uno de los tokens más usados por los usuarios (ver Figura 8).

Como puede observarse en la Tabla 9, y de forma similar al caso de los tokens del género de ficción, los términos usados en los encabezamientos se encontraban incluidos en las etiquetas, lo cual no siempre ocurrió de forma opuesta. Sin embargo, destacó el grado de coincidencia entre ambos conjuntos de datos con respecto a la frecuencia de uso para el término *biography*. También es de notar la cantidad de empates en el caso de los encabezamientos. A saber, *states* y *united*, *century* y *social*, así como *american* y *conditions*. Este fenómeno, el cual se debe a características propias de la descripción

Figura 8

Tokens en inglés más frecuentes, no ficción



bibliográfica como el uso de subdivisiones, explica el término adicional al final de la lista.

Similitud léxica. Como se mencionó, el coeficiente de Jaccard calcula la similitud entre dos conjuntos sin tomar en cuenta la frecuencia de sus elementos. Bajo esa premisa, y considerando cada conjunto de datos como un todo, el grado de similitud entre las etiquetas y los encabezamientos asociados al género de ficción fue bajo. Lo mismo ocurrió con la no ficción, aunque su coeficiente de Jaccard fue ligeramente más alto que el caso anterior (ver Tabla 10).

Nuevamente, considerando cada conjunto de datos como un todo, los valores del radio de cobertura para ambos grupos indicaron diferentes niveles de inclusión. Tal como los tokens más frecuentes dejaron entrever, los términos usados en los encabezamientos se encontraban incluidos dentro de las etiquetas en mayor medida que en el escenario opuesto. No obstante, más de la mitad de los términos usados en las etiquetas tuvieron

Tabla 9

Frecuencia de los tokens en inglés más frecuentes, no ficción

Encabezamientos			Etiquetas		
Tokens	Frecuencia	Frecuencia en etiquetas	Tokens	Frecuencia	Frecuencia en encabezamientos
states	250	20	*nonfiction*	384	1
united	250	20	*biography*	257	185
biography	185	257	*history*	226	48
politics	83	161	*memoir*	206	-
government	81	51	*usa*	170	-
history	48	226	*autobiography*	166	-
presidents	47	60	*politics*	161	83
century	43	90	*american*	154	34
social	43	44	*science*	97	7
american	34	154	*political*	93	28
conditions	34	3			

Tabla 10

Similitud de los tokens en inglés por conjunto de datos

Tipo de literatura	Coeficiente de Jaccard	Radio de cobertura de los encabezamientos	Radio de cobertura de las etiquetas
Ficción	0,3	0,8	0,5
No ficción	0,4	0,9	0,6

coincidencias con los encabezamientos, tanto en el caso de la ficción como el de la no ficción.

Tras analizar la similitud entre los encabezamientos y las etiquetas por cada título, se obtuvieron valores más bajos que al considerar a cada conjunto de datos como un todo, como era previsible. Es así como el promedio del coeficiente de Jaccard para los grupos de

ficción y no ficción correspondió a 0,2 en ambos casos, lo que indica que no hubo muchas coincidencias entre los términos que emplearon los usuarios y los catalogadores para cada título individual. Los conjuntos de datos de no ficción fueron también los que presentaron mayor variabilidad con un mínimo de 0 y un máximo de 0,6, frente al mínimo de 0 y el máximo de 0,4 de los conjuntos de datos de ficción (ver Figura 9).

En el caso de los conjuntos de datos de ficción, el promedio del radio de cobertura de los encabezamientos fue de 0,5, mientras que el promedio del radio de cobertura de las etiquetas fue de 0,2, lo que implica que la mitad de los términos asignados por los catalogadores coincidieron con los términos asignados por los usuarios. En contraposición, menos de un cuarto de los términos asignados por los usuarios coincidieron con los términos asignados por los catalogadores (ver Figura 10).

Figura 9

Coeficiente de Jaccard de los tokens en inglés por título

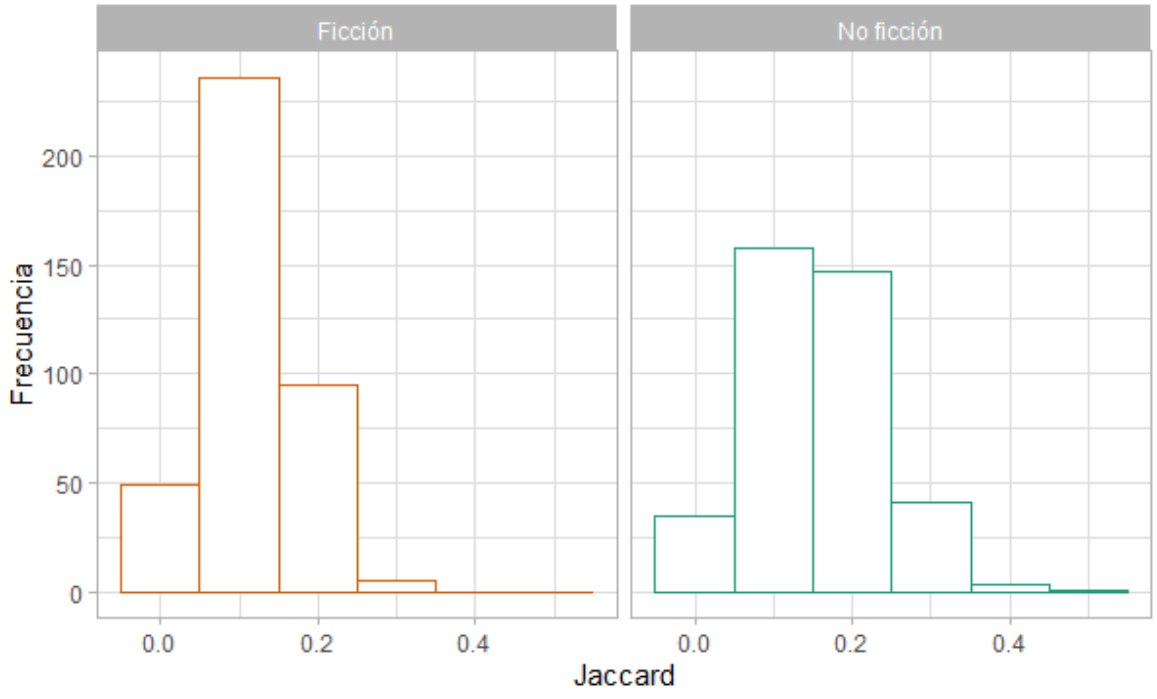
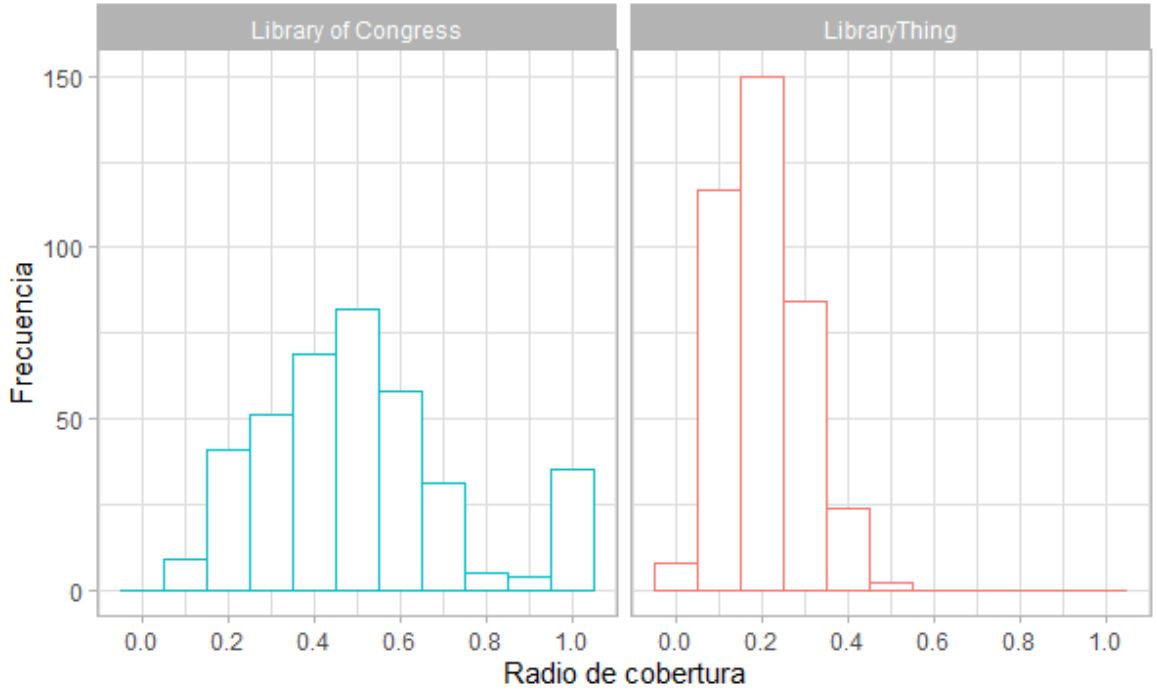


Figura 10

Radio de cobertura de los tokens en inglés por título, ficción

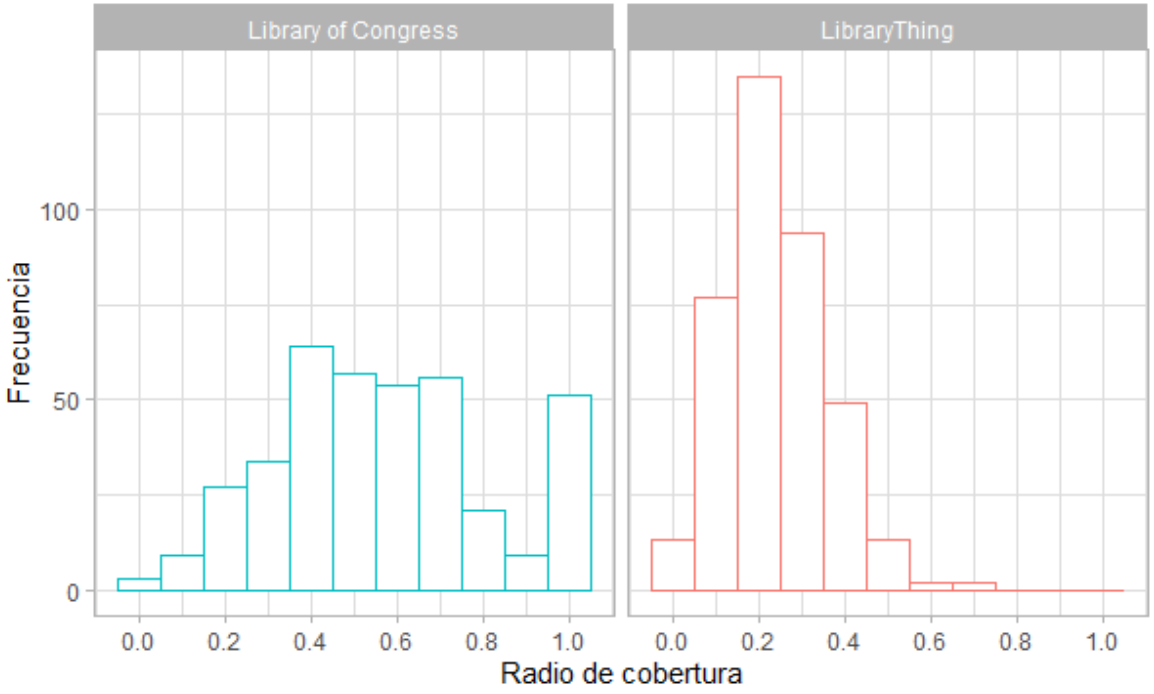


En el caso de los conjuntos de datos de no ficción, el promedio del radio de cobertura de los encabezamientos fue de 0,6, mientras que el promedio del radio de cobertura de las etiquetas fue de 0,2, lo que implica que más de la mitad de los términos asignados por los catalogadores coincidieron con los términos asignados por los usuarios. En contraste, cerca de un cuarto de los términos asignados por los usuarios coincidieron con los términos asignados por los catalogadores (ver Figura 11).

Asociación entre términos. De acuerdo con Silge y Robinson (2021), los términos que están cerca de la línea recta cuentan con frecuencias similares en ambos conjuntos, como fue el caso de *fiction*. Mientras tanto, aquellos que están lejos de la línea son términos que están presentes en un conjunto de datos en mayor medida que en el otro. Es así que *character* (personaje) resultó ser un término representativo del conjunto de datos de la Library of Congress. Algo similar ocurrió con el término *mystery* (misterio), hallado

Figura 11

Radio de cobertura de los tokens en inglés por título, no ficción



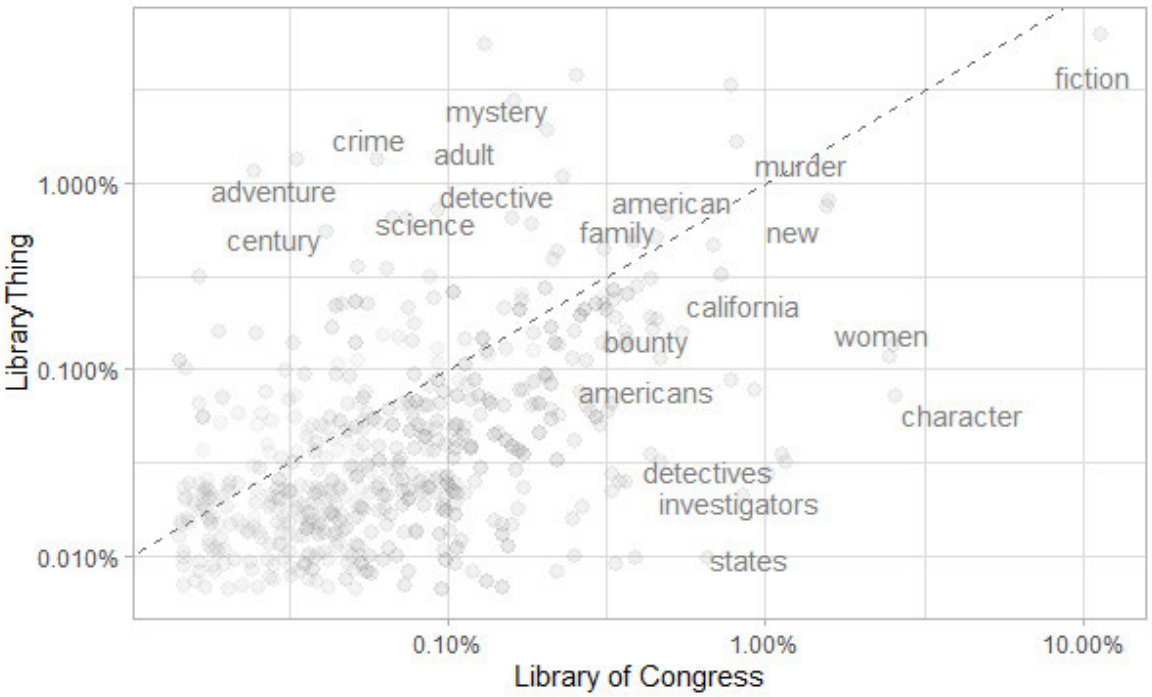
con más frecuencia en el conjunto de datos de LibraryThing y que de hecho tiene un significado distinto en el caso de los encabezamientos (ver Figura 12).

Silge y Robinson (2021) cuantifican la similitud entre las frecuencias de uso de los términos a través del coeficiente de Pearson. Sin embargo, en la presente investigación se optó por evaluar la correlación a través del coeficiente de Kendall, cuyo valor para los conjuntos de datos de ficción fue de 0,5.

Biography (biografía) fue uno de los términos con frecuencias similares dentro de los conjuntos de datos de no ficción. Lo mismo ocurrió con los términos *20th* y *21st* como parte de las subdivisiones cronológicas referentes a los siglos XX y XXI. En cuanto a los términos que se encontraban en un conjunto en mayor medida que en el otro, el término *states*, en referencia al nombre del país norteamericano, resultó ser más representativo del conjunto de datos de los encabezamientos. De igual manera, el término *business* (negocio,

Figura 12

Distribución de la frecuencia de los tokens en inglés, ficción



de acuerdo con la traducción provista por LibraryThing) se encontró con mayor frecuencia dentro de las etiquetas. Finalmente, el coeficiente de Kendall fue de 0,5 al igual que en el caso de los conjuntos de datos de ficción (ver Figura 13).

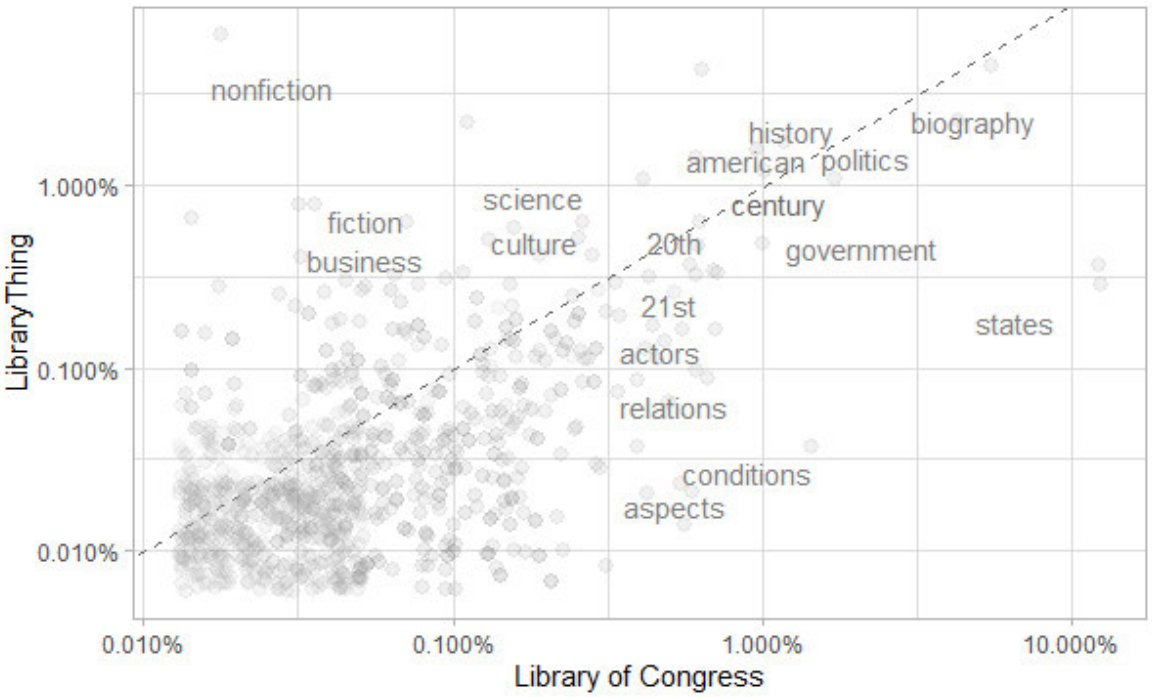
Etiquetas en español

Al igual que en el caso de sus pares en inglés, tanto el conjunto de datos de ficción como el de no ficción contaron con 11 550 y 11 536 etiquetas, respectivamente. En cuanto al número de etiquetas únicas, el conjunto de datos de ficción contó con 1705 etiquetas, mientras que el de no ficción contó con 3126 etiquetas.

Como puede observarse en las nubes de etiquetas (ver Figuras 14 y 15), los conjuntos de datos incluyeron ejemplos de falta de uniformidad en el uso de mayúsculas, términos ambiguos como abreviaturas y siglas, términos en otros idiomas como *american politics* (política estadounidense) o *large print* (letra grande), términos traducidos

Figura 13

Distribución de la frecuencia de los tokens en inglés, no ficción



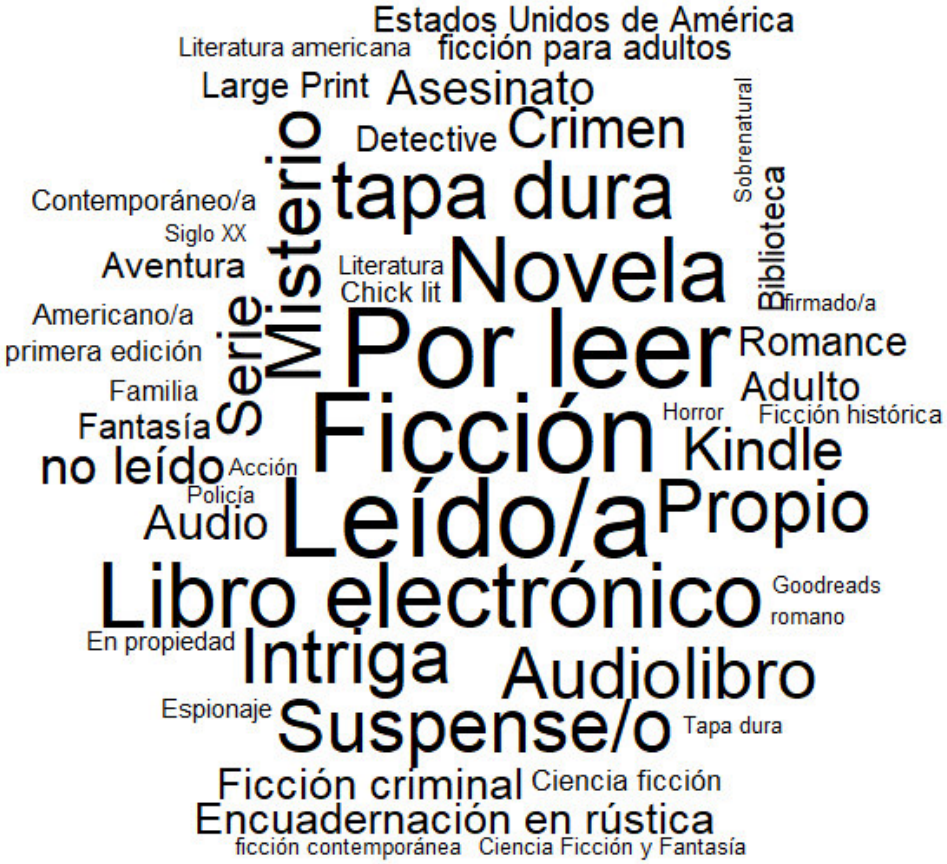
incorrectamente como “romano” para referirse a *roman* (novela, en alemán) o *suspense/o* para referirse a suspenso, sinónimos como “actualidad” y “temas de actualidad”, descripciones de acciones planeadas como “leído/a” y “por leer”, referencias acerca del formato como “libro electrónico” y “encuadernación en rústica”, así como referencias acerca del ejemplar como “primera edición” y “firmado/a”.

A pesar de las variaciones que se mencionaron con respecto a la recolección de datos, las etiquetas en español coincidieron con sus pares en inglés tanto en la frecuencia de uso como en la clase de términos que destacaron en cada nube de etiquetas. A saber, los géneros literarios en el caso de los títulos de ficción y los términos relacionados con temas de actualidad en el caso de los títulos de no ficción.

Como parte del preprocesamiento de datos, la eliminación de las etiquetas usadas por un solo usuario y aquellas que indicaban el precio de los ejemplares tuvo efectos

Figura 14

Etiquetas en español más frecuentes, ficción



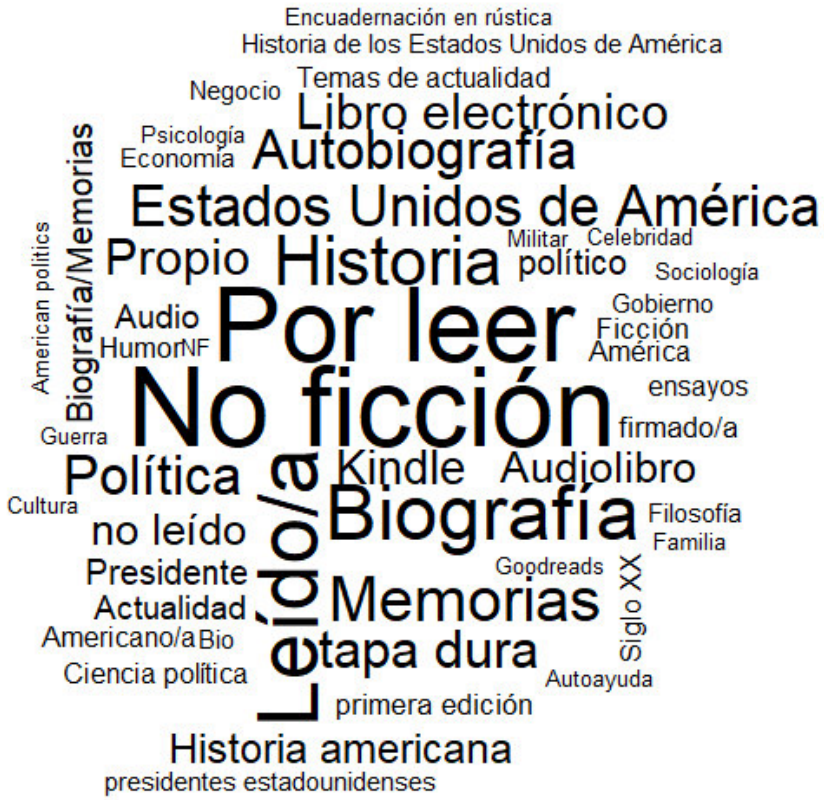
similares a los observados en los conjuntos de datos en inglés. En el caso de los títulos de ficción, el número total de etiquetas se redujo en 1,1%, mientras que las etiquetas únicas se vieron reducidas en 5,6%. En el caso de los títulos de no ficción, la disminución fue igual de notoria que la del conjunto de datos en inglés. Efectivamente, el número total de etiquetas se redujo en 9,4%, mientras que las etiquetas únicas se vieron reducidas en 27,7%. La misma explicación —la diversidad en términos de contenido— se aplica a este caso.

Encabezamientos en español

Tras la recolección de datos y la selección de los campos 600, 610, 611, 630, 650, 651 y 655, cuya procedencia fuera la Lista de encabezamientos de la Library of

Figura 15

Etiquetas en español más frecuentes, no ficción



Congress, el conjunto de datos de ficción estuvo conformado por 1493 encabezamientos complejos, de los cuales 789 resultaron ser únicos. Mientras tanto, el conjunto de datos de no ficción estuvo conformado por 1509 encabezamientos complejos, de los cuales 1103 resultaron ser únicos. A pesar de que inicialmente los términos de ficción fueron más numerosos, se confirmó la tendencia relacionada con la variabilidad de los términos asignados a los títulos de no ficción.

Tokens en español

Al igual que en el caso de los tokens en inglés, tras la tokenización, y continuando con el preprocesamiento de datos, la eliminación de las palabras vacías adicionales a las del paquete stopwords se dio en función a las etiquetas. Si bien la mayoría de términos

considerados como no significativos coincidió con las palabras vacías en inglés, existieron algunas variaciones en la frecuencia de uso.

Tal como puede observarse en la Tabla 11, el porcentaje en que se redujo el número de tokens fue mayor en aquellos provenientes de las etiquetas de ficción, mientras que lo opuesto ocurrió con los tokens provenientes de los encabezamientos de no ficción. Sin embargo, como puede observarse en la Tabla 12, y de manera similar a lo que ocurrió con los tokens únicos en inglés, esto no tuvo un efecto considerable en el número de tokens únicos en español.

Como resultado del preprocesamiento, el promedio de tokens por título fue de 22,7 y 24,1 en el caso de las etiquetas de ficción y no ficción, respectivamente, mientras que en el caso de los encabezamientos de ficción y no ficción, el promedio de tokens por título fue de 8,9 y 9,3, respectivamente.

Frecuencia. En cuanto a los tokens más frecuentes en los conjuntos de datos de ficción, el hecho de que el catálogo de autoridades LIBRUNAM haya designado a "novela" como traducción de la subdivisión de forma *fiction* ocasionó que se perdiera la coincidencia entre ambos términos. A pesar de ello, se mantuvo la coincidencia entre la

Tabla 11

Tokens en español antes y después del preprocesamiento

Tipo de literatura	Antes	Después	Disminución
Etiquetas			
Ficción	17520	8722	50,2%
No ficción	15850	9258	41,6%
Encabezamientos			
Ficción	5482	3421	37,6%
No ficción	6911	3570	48,3%

Tabla 12

Tokens únicos en español antes y después del preprocesamiento

Tipo de literatura	Antes	Después	Disminución
Etiquetas			
Ficción	1740	1553	10,8%
No ficción	2408	2205	8,4%
Encabezamientos			
Ficción	1016	941	7,4%
No ficción	1434	1259	12,2%

traducción y la siguiente etiqueta más frecuente en el conjunto de datos de LibraryThing (ver Figura 16).

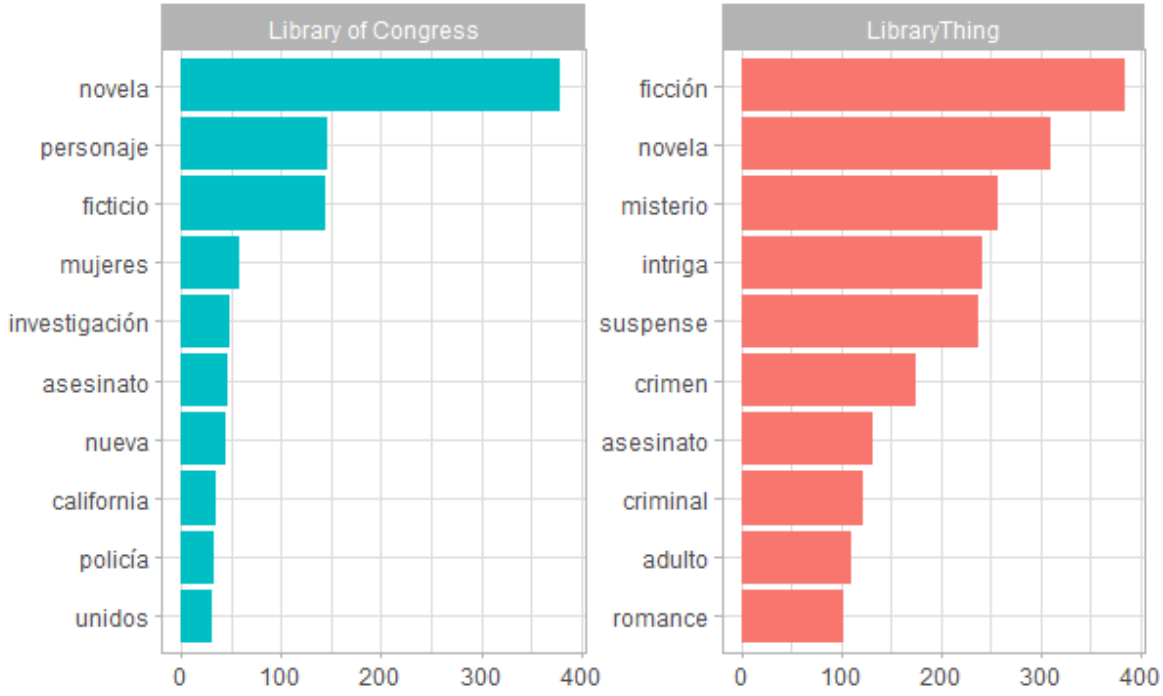
Al igual que en el caso de los tokens en inglés, el calificativo “personaje ficticio” apareció con frecuencia. De la misma manera, ambos conjuntos de datos coincidieron con sus pares en inglés al centrarse en géneros literarios, en especial aquellos de corte policial. Adicionalmente, los tokens provenientes de los encabezamientos incluyeron algunos topónimos que originalmente se emplearon en conjunción con la subdivisión de forma “novela”, siguiendo el mismo patrón de sus pares en inglés al corresponder a nombres corporativos, nombres geográficos y subdivisiones geográficas.

Como puede observarse en la Tabla 13, y a diferencia de lo que ocurrió con los tokens en inglés, tanto el conjunto de datos de la Library of Congress como el de LibraryThing contaban con términos que no aparecieron en el conjunto opuesto. Otra diferencia notoria fue la presencia de un término en inglés, *suspense*, originalmente *suspense/o*, y que fue considerado anteriormente como un error de traducción por parte de LibraryThing.

Entre las similitudes con los tokens en inglés puede mencionarse el grado de coincidencia entre ambos conjuntos de datos con respecto al término “novela”. A pesar de

Figura 16

Tokens en español más frecuentes, ficción



que en este caso la concordancia se dio entre el token más frecuente del conjunto de datos de los encabezamientos y el segundo token más frecuente del conjunto de datos de las etiquetas, ambas cantidades son similares en comparación con la frecuencia de uso del resto de tokens.

Con respecto a los conjuntos de datos de no ficción, el término más usado en los encabezamientos correspondió a una parte del topónimo “Estados Unidos”, encontrado también en el conjunto de datos opuesto. Debido a que los glosarios en español de stopwords consideran a “estados” como una palabra vacía, tanto los encabezamientos como las etiquetas carecieron de este término luego del preprocesamiento (ver Figura 17).

Al igual que en el caso anterior, el término “no ficción” fue el más frecuente dentro del grupo de las etiquetas. Asimismo, los dos conjuntos de datos coincidieron en el uso de los términos “biografía”, “historia” y “política”, aunque a diferencia de los tokens en

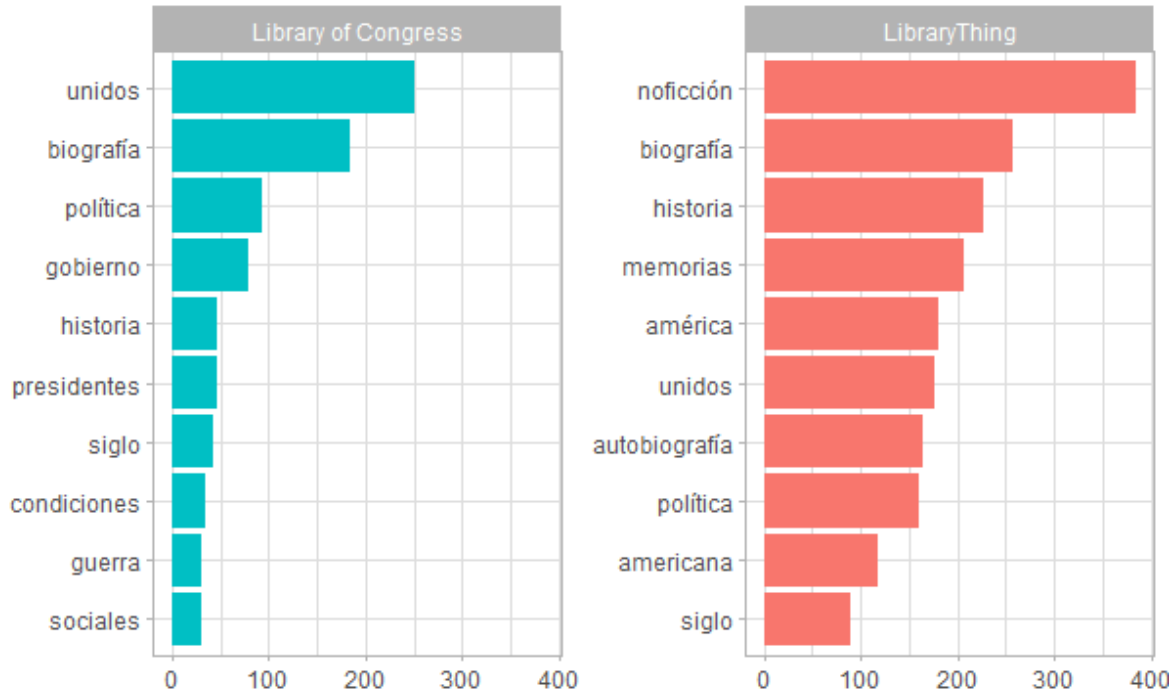
Tabla 13

Frecuencia de los tokens en español más frecuentes, ficción

Encabezamientos			Etiquetas		
Tokens	Frecuencia	Frecuencia en etiquetas	Tokens	Frecuencia	Frecuencia en encabezamientos
novela	379	310	ficción	385	3
personaje	147	-	novela	310	379
ficticio	145	-	misterio	257	5
mujeres	58	12	intriga	242	-
investigación	49	-	*suspense*	238	-
asesinato	47	131	crimen	175	2
nueva	44	51	asesinato	131	47
california	35	23	criminal	122	9
policía	32	40	adulto	111	-
unidos	30	81	romance	103	-

Figura 17

Tokens en español más frecuentes, no ficción



inglés, el término “siglo” también apareció entre los tokens más frecuentes en ambos grupos, desplazando al término “ciencia” en el caso de las etiquetas.

Tal como se advirtió en el caso de los tokens del género de ficción, ambos conjuntos de datos contaban con términos que no aparecieron en el conjunto opuesto. En el caso de los encabezamientos, esto ocurrió con “condiciones”, término usado comúnmente durante la descripción mediante subdivisiones generales como “condiciones económicas” o “condiciones sociales” (ver Tabla 14). Entre las similitudes con los tokens en inglés puede mencionarse el grado de coincidencia entre ambos conjuntos de datos con respecto al término “biografía”. Si bien las frecuencias de uso varían entre sí, la brecha fue mucho menor que en el caso de los otros tokens.

Similitud léxica. Considerando cada conjunto de datos como un todo, el grado de similitud entre las etiquetas y los encabezamientos asociados al género de ficción fue bajo, incluso más que en el caso de los tokens en inglés. Lo mismo ocurrió con la no ficción, aunque en menor medida (ver Tabla 15).

En cuanto al radio de cobertura de los encabezamientos, los valores para la ficción y la no ficción indicaron que más de la mitad de los términos se encontraban incluidos en el conjunto de datos de las etiquetas. Si bien el radio de cobertura de las etiquetas fue menor, la diferencia frente al radio de cobertura de los encabezamientos no fue tan marcada como ocurrió con los tokens en inglés.

Con respecto a la similitud entre los encabezamientos y las etiquetas por cada título, el promedio del coeficiente de Jaccard para los grupos de ficción y no ficción fue igual a 0,1 y 0,2, respectivamente, de forma similar a los coeficientes de los tokens en inglés. En general, ello indica que las coincidencias entre los términos que emplearon los usuarios y los catalogadores para cada título individual tampoco fueron numerosas en el caso de los términos en español (ver Figura 18).

Tabla 14

Frecuencia de los tokens en español más frecuentes, no ficción

Encabezamientos			Etiquetas		
Tokens	Frecuencia	Frecuencia en etiquetas	Tokens	Frecuencia	Frecuencia en encabezamientos
unidos	251	177	nofícción	384	-
biografía	185	258	biografía	258	185
política	93	160	historia	227	47
gobierno	80	51	memorias	207	2
historia	47	227	américa	181	-
presidentes	47	45	unidos	177	251
siglo	43	90	autobiografía	165	-
condiciones	34	-	política	160	93
guerra	31	57	americana	118	-
sociales	31	22	siglo	90	43

Tabla 15

Similitud de los tokens en español por conjunto de datos

Tipo de literatura	Coeficiente de similitud de Jaccard	Radio de cobertura de los encabezamientos	Radio de cobertura de las etiquetas
Ficción	0,2	0,6	0,5
No ficción	0,3	0,7	0,5

En el caso de los conjuntos de datos de ficción, el promedio del radio de cobertura de los encabezamientos fue de 0,4, valor ligeramente menor al obtenido en el caso de los tokens en inglés. Por otro lado, el promedio del radio de cobertura de las etiquetas fue de 0,2, al igual que en el caso de los tokens en inglés (ver Figura 19).

En el caso de los conjuntos de datos de no ficción, el promedio del radio de cobertura de los encabezamientos fue de 0,5, mientras que el promedio del radio de

Figura 18

Coeficiente de Jaccard de los tokens en español por título

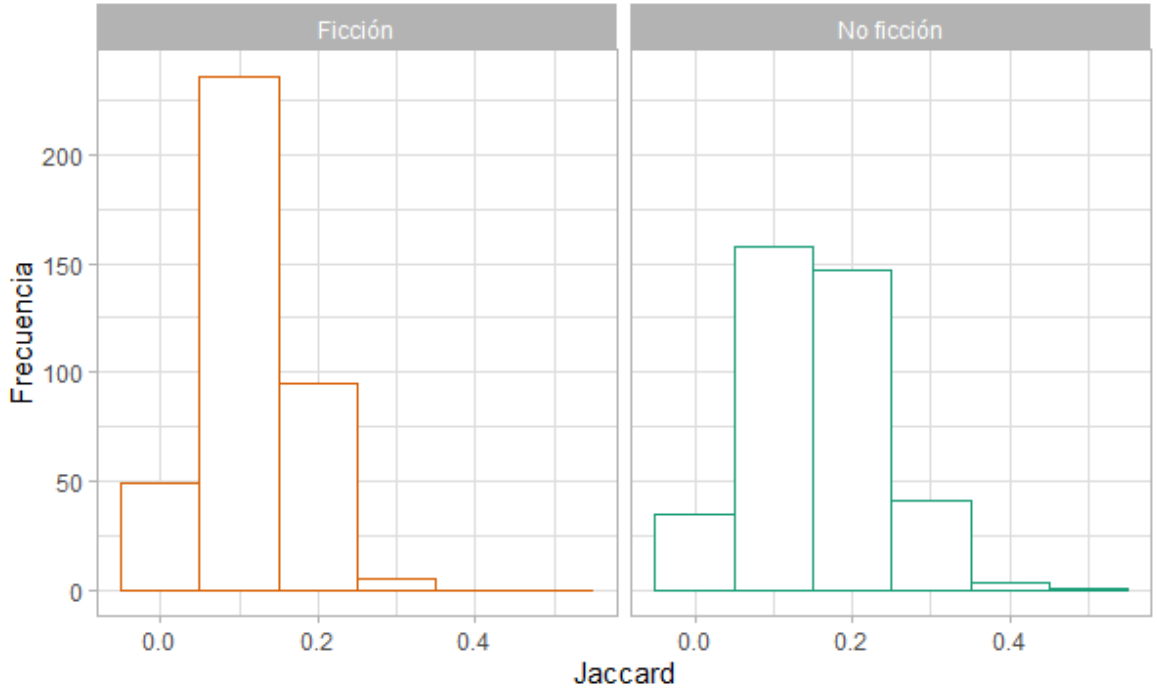
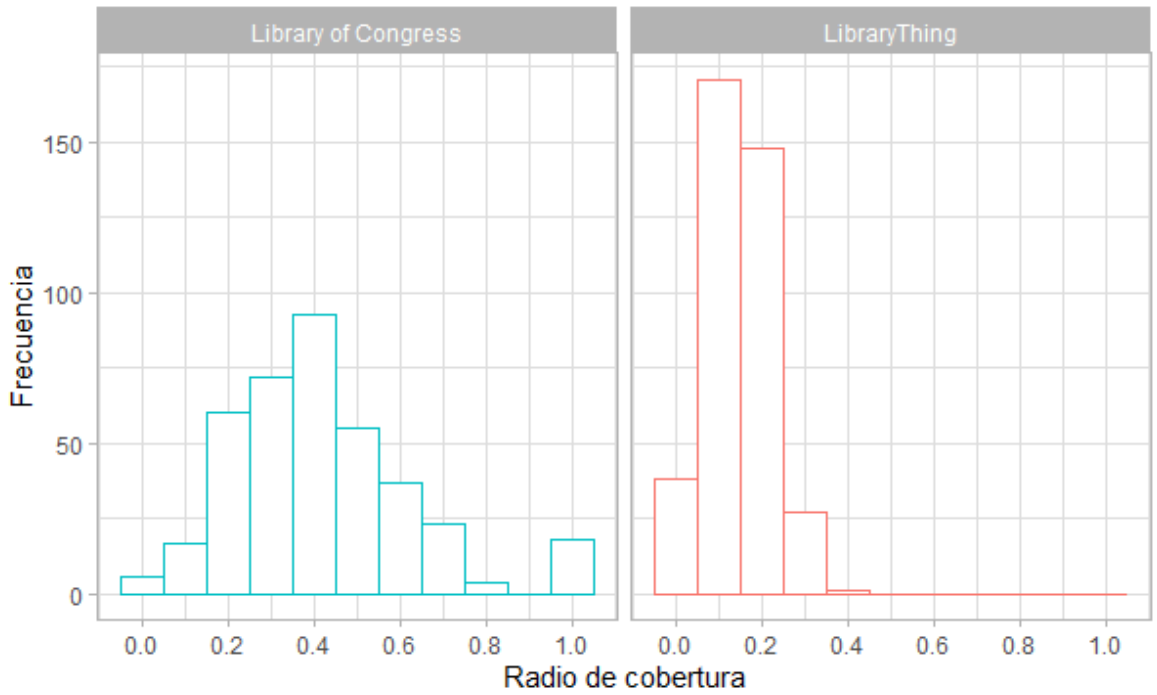


Figura 19

Radio de cobertura de los tokens en español por título, ficción



cobertura de las etiquetas fue de 0,2. El primer valor fue ligeramente menor al obtenido en el caso de los tokens en inglés, pero el segundo de ellos fue idéntico (ver Figura 20).

Asociación entre términos. "Historia" fue uno de los términos con frecuencias similares dentro de los conjuntos de datos de ficción, mientras que "personas" y "ficción" sirvieron como ejemplos de términos representativos de los conjuntos de datos de la Library of Congress y LibraryThing, respectivamente. Al igual que en el caso de los tokens en inglés, el coeficiente de Kendall fue de 0,5 (ver Figura 21).

En cuanto a los conjuntos de datos de no ficción, "estadounidense" fue uno de los términos con frecuencias similares en ambos conjuntos, mientras que "unidos" y "memorias" sirvieron como ejemplos de términos representativos de los conjuntos de datos de la Library of Congress y LibraryThing, respectivamente. Al igual que en el caso de los conjuntos de datos de ficción, el coeficiente de Kendall fue de 0,5 (ver Figura 22).

Figura 20

Radio de cobertura de los tokens en español por título, no ficción

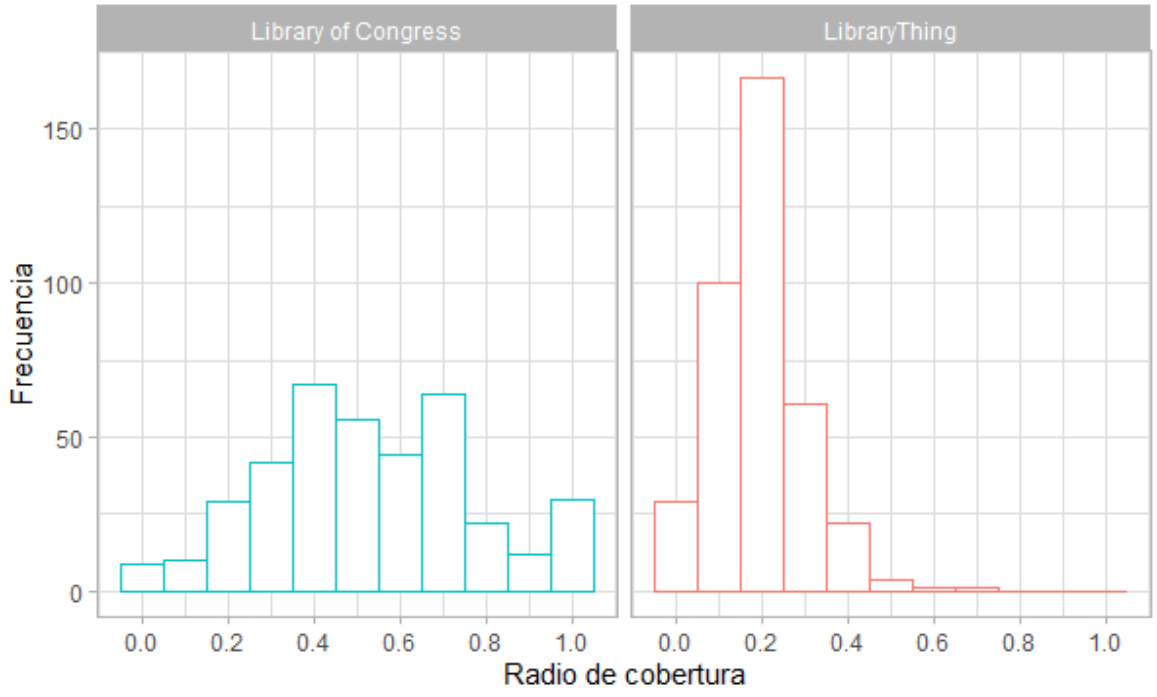


Figura 21

Distribución de la frecuencia de los tokens en español, ficción

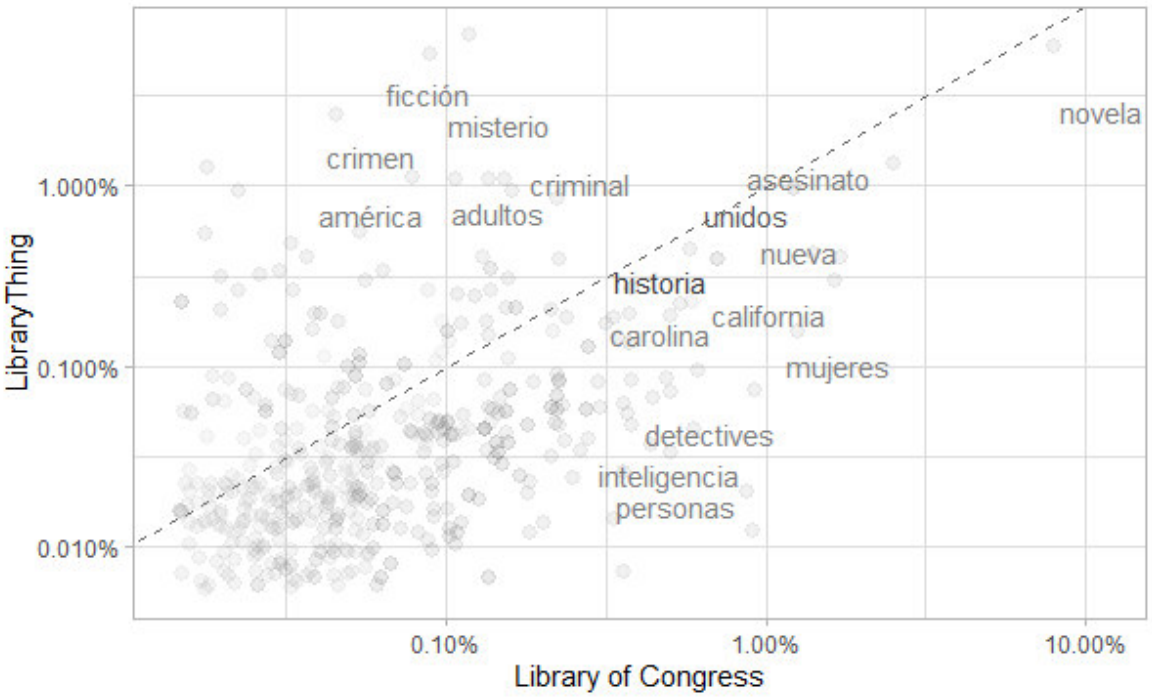
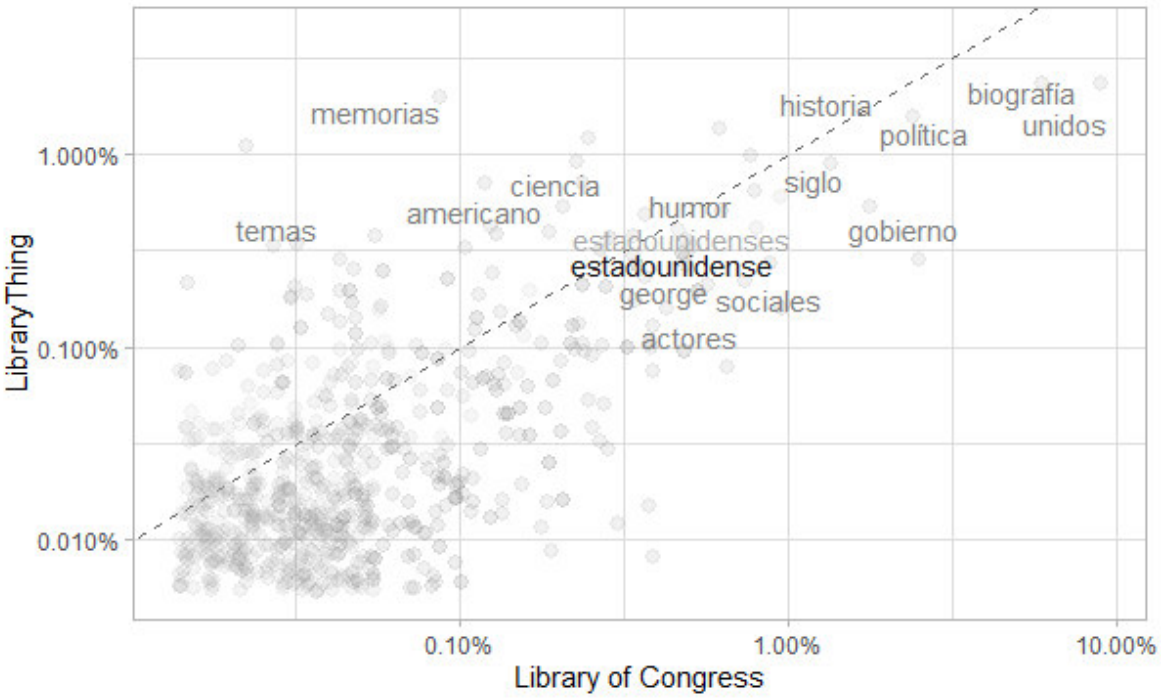


Figura 22

Distribución de la frecuencia de los tokens en español, no ficción



Capítulo VI: Conclusiones y recomendaciones

Conclusiones

1. Las folksonomías, colecciones de términos empleados por los usuarios de un servicio para describir temáticamente recursos de información mediante el uso del lenguaje natural, forman parte del entorno de comunidades virtuales de catalogación, tales como LibraryThing en el caso de libros. A diferencia de vocabularios controlados como los encabezamientos de materia de la Library of Congress, las folksonomías carecen de relaciones jerárquicas y mecanismos de control exhaustivos que les permitan convertirse en sustitutos de los primeros, lo que no impide que puedan considerarse como un complemento en la descripción de contenido.
2. Los patrones observados en la asignación de términos por parte de los usuarios incluyeron palabras vacías, términos ambiguos debido a razones que iban desde la escritura hasta la polisemia, así como términos que no hacían referencia al contenido de los libros de forma directa, reflejando la problemática asociada con las folksonomías. Si esto pudo observarse en una comunidad de usuarios como la de LibraryThing, con mayor razón ocurrirá en el caso de servicios que no cuenten con una comunidad lo suficientemente numerosa y activa como para que el proceso de estabilización semántica emerja de forma natural.
3. Se asumió que el grado de similitud entre los términos sería bajo considerando los resultados de estudios anteriores. Efectivamente, el coeficiente de Jaccard fue bajo en todos los casos, lo que significa que los términos coincidentes no fueron considerables. Al mismo tiempo, el radio de cobertura de los encabezamientos superó en todos los casos al de las etiquetas, lo que se explica gracias a que los

términos asignados por los usuarios suelen ser más numerosos y, por tanto, incluyen a los asignados por los catalogadores en mayor medida.

4. Asimismo, se asumió que el grado de asociación entre los términos coincidentes sería moderado considerando los resultados de estudios anteriores. En efecto, la correlación expresada a través del coeficiente de Kendall fue moderada para todos los conjuntos de datos, lo que implica que tanto los usuarios como los catalogadores asignaron una prioridad similar a los términos coincidentes en la mitad de los casos. De ello se desprende también que los encabezamientos no se encuentran alejados del uso común del lenguaje a pesar de contar con un nivel mayor de rigurosidad.
5. En general, la comparación entre ambos conjuntos de términos reveló que las etiquetas sociales son lo suficientemente distintas de los encabezamientos de materia como para justificar su inclusión dentro de la descripción. Su complementariedad se debe a elementos como la flexibilidad, la acuñación de neologismos y el grado de especificidad al que pueden llegar las etiquetas, lo cual, aunado a la estructura y la contextualización terminológica que brindan los encabezamientos, incrementa la accesibilidad de la información y facilita la exploración de recursos relacionados que sean de interés para los usuarios.

Recomendaciones

1. Hoy en día existen diversos sistemas integrados de gestión bibliotecaria que contemplan la asignación de etiquetas, permitiendo que los usuarios organicen los recursos de información a través de términos definidos por ellos mismos. En nuestro medio, las bibliotecas académicas son probablemente las que podrían implementar dicha funcionalidad con mayor facilidad gracias al presupuesto que

tienen designado para la adquisición de software bibliotecario, sin mencionar que su relación directa con sus usuarios favorecería la promoción del servicio.

2. Considerando la presencia de etiquetas idiosincráticas, podría establecerse un control mínimo antes de visibilizar las etiquetas asignadas a un recurso de información. Asimismo, podría recurrirse a fuentes externas que enriquezcan los registros bibliográficos a través de metadatos complementarios, sirviendo como referencia para los usuarios con respecto al tipo de términos que pueden emplear para la descripción.
3. Las coincidencias entre el vocabulario empleado por los usuarios y los catalogadores pueden dar lugar a proyectos de *crowdsourcing*. Una vez seleccionados los recursos cuyos metadatos se busca enriquecer, podría fomentarse la participación de los usuarios en la descripción y transcripción de los contenidos, logrando tanto su acercamiento a las colecciones de la biblioteca como la construcción de conocimiento de utilidad para la comunidad.
4. La relevancia de la ciencia de datos para la bibliotecología no se limita a la minería de textos, ya que también puede permitir a los usuarios interactuar con información específica a través de técnicas de visualización de datos. Asimismo, campos relacionados como el aprendizaje automático pueden servir para agrupar imágenes similares en colecciones de materiales gráficos que no cuenten con datos suficientes para su correcta identificación.
5. Futuras líneas de investigación podrían incluir en la comparación a registros de autoridades provenientes de unidades de información más cercanas a nuestra realidad como la Biblioteca Nacional de Brasil. Asimismo, se podría considerar no solamente a monografías sino también a materiales especiales cuyas características podrían requerir de un nivel adicional de representación del contenido. Por último,

se podría ir más allá de la similitud léxica para enfocarse en la similitud semántica y la efectividad en la recuperación de la información, de modo que pueda determinarse el resultado de la interacción entre las etiquetas sociales y los encabezamientos de materia.

Referencias bibliográficas

Aghaebrahimian, A., Stauder, A., & Ustaszewski, M. (2020). Testing the validity of Wikipedia categories for subject matter labelling of open-domain corpus data. *Journal of Information Science*. https://doi.org/10.1177/0165551520977438

Anderson, P. (2012). *Web 2.0 and beyond: principles and technologies* [La web 2.0 y más allá: principios y tecnologías]. CRC Press. https://books.google.com.pe/books?id=rRrOBQAAQBAJ

Argimon Pallás, J. M., & Jiménez Villa, J. (2019). *Métodos de investigación clínica y epidemiológica* (5.ª ed.). Elsevier. https://books.google.com.pe/books?id=ogCiDwAAQBAJ

Barros, L. M. S. (2011). *A folksonomia como prática de classificação colaborativa para a recuperação da informação* [Tesis de maestría, Universidade Federal do Rio de Janeiro]. http://ridi.ibict.br/handle/123456789/737

Batley, S. (2014). *Classification in theory and practice* [La clasificación en la teoría y en la práctica] (2.ª ed.). Chandos Publishing. https://books.google.com.pe/books?id=yryUAwAAQBAJ

Belém, F. M., Almeida, J. M., & Gonçalves, M. A. (2017). A survey on tag recommendation methods. *Journal of the Association for Information Science and Technology*, *68*(4), 830-844. https://doi.org/10.1002/asi.23736

Benoit, K., Muhr, D., & Watanabe, K. (2021). *stopwords: multilingual stopword lists*. The Comprehensive R Archive Network. https://cran.r-project.org/package=stopwords

Bernico, M. (2018). *Deep learning quick reference: useful hacks for training and optimizing deep neural networks with TensorFlow and Keras* [Referencia rápida sobre aprendizaje profundo: trucos útiles para entrenar y optimizar redes neuronales

profundas con TensorFlow y Keras]. Packt. https://books.google.com.pe/books?id=M5RRDwAAQBAJ

Bogers, T., & Petras, V. (2017). Supporting book search: a comprehensive comparison of tags vs. controlled vocabulary metadata. *Data and Information Management*, *1*(1), 17-34. https://doi.org/10.1515/dim-2017-0004

Brandt, M., & Medeiros, M. B. B. (2010). Folksonomia: esquema de representação do conhecimento? *Transinformação*, *22*(2), 111-121. https://www.scielo.br/j/tinf/a/F8mxgMCbfMYTjYvCXpPQtgd

Bullard, J. (2019). Curated folksonomies: three implementations of structure through human judgment. *Knowledge Organization*, *45*(8), 643-652. http://hdl.handle.net/2429/72779

Caballero González, C. (2016). *Pruebas de funcionalidades y optimización de páginas web: UF1306*. Paraninfo. https://books.google.com.pe/books?id=JH-mCwAAQBAJ

Campesato, O. (2021). *Natural language processing fundamentals for developers* [Fundamentos del procesamiento de lenguaje natural para desarrolladores]. Mercury Learning and Information. https://books.google.com.pe/books?id=Do00EAAAQBAJ

Chowdhury, G. G., Burton, P. F., McMenemy, D., & Poulter, A. (2008). *Librarianship: an introduction* [Bibliotecología: una introducción]. Facet Publishing. https://books.google.com.pe/books?id=idoqDgAAQBAJ

Cole, T. W., & Han, M-J. K. (2013). *XML for catalogers and metadata librarians* [XML para catalogadores y bibliotecarios de metadatos]. Libraries Unlimited. https://books.google.com.pe/books?id=DYIQAgAAQBAJ

Contreras Barrera, M. (2014). Minería de texto: una visión actual. *Biblioteca Universitaria, 17*(2), 129-138. https://www.redalyc.org/articulo.oa?id=28540279005

Dattolo, A., Ferrara, F., & Tasso, C. (2012). On social semantic relations for recommending tags and resources using folksonomies. En Z. S. Hippe, J. L. Kulikowski, & T. Mroczek (Eds.), *Human-computer systems interaction: backgrounds and applications 2* (Part 1, pp. 311-326). Springer. https://books.google.com.pe/books?id=UeQ8kR6lxXMC

De Meo, P., Ferrara, E., Abel, F., Aroyo, L., & Houben, G-J. (2013). Analyzing user behavior across social sharing environments. *ACM Transactions on Intelligent Systems and Technology*, *5*(1). https://arxiv.org/abs/1310.4399

Derham, R., & Mills, A. (2010). Web 2.0—social bookmarking: an overview of folksonomies. En S. Murugesan (Ed.), *Handbook of research on web 2.0, 3.0, and X.0: technologies, business, and social applications* (Vol. 1, pp. 206-224). Information Science Reference. https://books.google.com.pe/books?id=2LI9AQN1HIcC

DeZelar-Tiedman, C. (2011). Exploring user-contributed metadata's potential to enhance access to literary works: social tagging in academic library catalogs. *Library Resources & Technical Services*, *55*(4), 221-233. https://doi.org/10.5860/lrts.55n4.221

Dong, H., Wang, W., & Coenen, F. (2018). Learning relations from social tagging data. En X. Geng, & B-H. Kang (Eds.), *PRICAI 2018: trends in artificial intelligence: 15th Pacific Rim International Conference on Artificial Intelligence, Nanjing, China, August 28–31, 2018, proceedings* (Part 1, pp. 29-41). Springer. https://books.google.com.pe/books?id=McpmDwAAQBAJ

Dorman, M. (2020). *Introduction to web mapping* [Introducción al mapeo web]. CRC Press. https://books.google.com.pe/books?id=Q3jNDwAAQBAJ

Feliciati, P. (2022). Call me by your name: towards an authority data control shared between archives and libraries. *JLIS.It*, *13*(1), 203-214. https://doi.org/10.4403/jlis.it-12733

Fellows, I. (2018). *wordcloud: word clouds*. The Comprehensive R Archive Network. https://cran.r-project.org/package=wordcloud

Fernández Ramos, A. (2017). Consumidores de información y descripción de documentos: las etiquetas de los usuarios en los catálogos de bibliotecas. En A. A. Rodríguez García, & R. A. González Castillo (Coords.), *Tendencias multidisciplinarias del uso de los metadatos* (pp. 117-128). Universidad Nacional Autónoma de México. http://dx.doi.org/10.22201/iibi.9786070299469e.2018

Gedikli, F. (2013). *Recommender systems and the social web: leveraging tagging data for recommender systems* [Los sistemas de recomendación y la web social: aprovechando los datos de etiquetado para los sistemas de recomendación]. Springer Wieveg. https://books.google.com.pe/books?id=YWVEAAAAQBAJ

Gilton, D. L. (2016). *Creating and promoting lifelong learning in public libraries: tools and tips for practitioners* [Creando y promoviendo el aprendizaje permanente en las bibliotecas públicas: herramientas y consejos para los profesionales]. Rowman & Littlefield. https://books.google.com.pe/books?id=_SxIDAAAQBAJ

Gómez-Díaz, R. (2013). *Etiquetar en la web social*. Editorial UOC. https://books.google.com.pe/books?id=faHGAgAAQBAJ

Guimarães, J. A. C., Moraes, J. B. E., & Guarido, M. D. M. (2007). Análisis documental de contenido de textos narrativos: bases epistemológicas y perspectivas

metodológicas. *Ibersid, Revista de Sistemas de Información y Documentación*, *1*, 93-99. https://doi.org/10.54886/ibersid.v1i.3267

Gupta, M., Li, R., Yin, Z., & Han, J. (2011). An overview of social tagging and applications. En C. C. Aggarwal (Ed.), *Social network data analytics* (pp. 447-492). Springer. https://books.google.com.pe/books?id=SE2iRgeYYwcC

Halpin, H. (2013). *Social semantics: the search for meaning on the web* [Semántica social: la búsqueda de significado en la web]. Springer. https://books.google.com.pe/books?id=o_dfm8NoIgYC

Hawes Publications. (2021, 10 de agosto). *New York Times adult hardcover best seller number ones listing*. Hawes Publications. http://www.hawes.com/number1s.htm

Heymann, P., & Garcia-Molina, H. (2009, 9-13 de febrero). *Contrasting controlled vocabulary and tagging: do experts choose the right names to label the wrong things?* [Ponencia]. Second ACM International Conference on Web Search and Data Mining, Barcelona, España. http://ilpubs.stanford.edu:8090/955

Hinton, P. R., McMurray, I., & Brownlow, C. (2014). *SPSS explained* [SPSS explicado] (2.ª ed.). Routledge. https://books.google.com.pe/books?id=Q_gjAwAAQBAJ

Hornik, K. (2020, 20 de febrero). *R FAQ: frequently asked questions on R*. The Comprehensive R Archive Network. https://cran.r-project.org/doc/FAQ/R-FAQ.html

Hjørland, B. (2017). Subject (of documents). *Knowledge Organization*, *44*(1), 55-64. https://doi.org/10.5771/0943-7444-2017-1-55

Iglesias Rebollo, C., & González Gordon, M. (2005). *Diccionario de propiedad intelectual: español / inglés / español*. Editorial Reus. https://books.google.com.pe/books?id=FsFUDwAAQBAJ

Ignatow, G., & Mihalcea, R. (2017). *Text mining: a guidebook for the social sciences* [Minería de textos: una guía para las ciencias sociales]. SAGE Publications. https://books.google.com.pe/books?id=EX6zDAAAQBAJ

Jacobson, S. (2014). Folksonomy. En K. Harvey (Ed.), *Encyclopedia of social media and politics* (Vol. 1, pp. 528-529). SAGE Reference. https://books.google.com.pe/books?id=qS91AwAAQBAJ

Johansson, S., & Golub, K. (2019). LibraryThing for libraries: how tag moderation and size limitations affect tag clouds. *Knowledge Organization*, *46*(4), 245-259. http://urn.kb.se/resolve?urn=urn:nbn:se:lnu:diva-88089

Kipp, M. E. I., Beak, J., & Choi, I. (2017). Motivations and intentions of Flickr users in enriching flick records for Library of Congress photos. *Journal of the Association for Information Science and Technology*, *68*(10), 2364-2379. https://doi.org/10.1002/asi.23869

Klašnja-Milićević, A., Vesin, B., Ivanović, M., Budimac, Z., & Jain, L. C. (2017). *E-learning systems: intelligent techniques for personalization* [Sistemas de aprendizaje virtual: técnicas inteligentes para la personalización]. Springer. https://books.google.com.pe/books?id=52-zDAAAQBAJ

Kopeinik, S., Lex, E., Seitlinger, P., Albert, D., & Ley, T. (2017). Supporting collaborative learning with tag recommendations: a real-world study in an inquiry-based classroom project. En *LAK '17: proceedings of the Seventh International Learning Analytics & Knowledge Conference* (pp. 409-418). ACM. https://doi.org/10.1145/3027385.3027421

Kumar, A., & Paul, A. (2016). *Mastering text mining with R: master text-taming techniques and build effective text-processing applications with R* [Dominando la minería de textos con R: vuélvete un experto en técnicas para dominar textos y crea

aplicaciones de procesamiento de texto efectivas con R]. Packt. https://books.google.com.pe/books?id=YNDcDgAAQBAJ

Laffly, D. (2020). Sampling strategies. En D. Laffly (Ed.), *TORUS 1 – toward an open resource using services: cloud computing for environmental data* (pp. 7-18). ISTE; Wiley. https://books.google.com.pe/books?id=E9XkDwAAQBAJ

Lee, D. H., & Schleyer, T. (2012). Social tagging is no substitute for controlled indexing: a comparison of Medical Subject Headings and CiteULike tags assigned to 231,388 papers. *Journal of the American Society for Information Science and Technology*, *63*(9), 1747-1757. https://doi.org/10.1002/asi.22653

Li, W., Grakova, N., & Qian, L. (2020). Ontological approach for question generation and knowledge control. En V. Golenkov, V. Krasnoproshin, V. Golovko, & E. Azarov (Eds.), *Open Semantic Technologies for Intelligent System: 10th International Conference, OSTIS 2020, Minsk, Belarus, February 19-22, 2020, revised selected papers* (pp. 161-175). Springer. https://books.google.com.pe/books?id=pfcEEAAAQBAJ

Marinho, L. B., Hotho, A., Jäschke, R., Nanopoulos, A., Rendle, S., Schmidt-Thieme, L., Stumme, G., & Symeonidis, P. (2012). *Recommender systems for social tagging systems* [Sistemas de recomendación para sistemas de etiquetado social]. Springer. https://books.google.com.pe/books?id=kP6180ETYVQC

Martínez Tamayo, A. M., & Mendes. P. V. (2015). *Diseño y desarrollo de tesauros*. Universidad Nacional de La Plata. https://libros.fahce.unlp.edu.ar/index.php/libros/catalog/book/68

Meza Vega, E. (2018). *Apuntes de algoritmia*. Editorial Universidad del Cauca. https://books.google.com.pe/books?id=FIewDwAAQBAJ

Michel, J-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., The Google Books Team, Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, N., Orwant, J., Pinker, S., Nowak, M. A., & Aiden, E. L. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, *331*(6014), 176-182. https://doi.org/10.1126/science.1199644

Mitchell, R. (2015). *Web scraping with Python: collecting data from the modern web* [Web scraping con Python: recopilando datos de la web moderna]. O'Reilly. https://books.google.com.pe/books?id=7z_fCQAAQBAJ

Mochón-Bezares, G., Méndez-Rodríguez, E., & Sorli-Rojo, A. (2017). Etiquetado social y blog-scraping como alternativa para la actualización de vocabularios controlados: aplicación práctica a un tesauro de biblioteconomía y documentación. *Información, cultura y sociedad*, *37*, 13-26. http://repositorio.filo.uba.ar/handle/filodigital/11218

Morley, D., & Parker, C. S. (2012). *Understanding computers: today and tomorrow* [Entendiendo a las computadoras: hoy y mañana] (14.ª ed.). Cengage Learning. https://books.google.com.pe/books?id=6cEKAAAAQBAJ

Morville, P., & Rosenfeld, L. (2006). *Information architecture for the World Wide Web* [Arquitectura de la información para la red informática mundial] (3.ª ed.). O'Reilly. https://books.google.com.pe/books?id=2d2Ry2hZc2MC

Mullen, L. (2020). *textreuse: detect text reuse and document similarity*. The Comprehensive R Archive Network. https://cran.r-project.org/package=textreuse

Müller, A., & Guido, S. (2017). *Introduction to machine learning with Python: a guide for data scientists* [Introducción al aprendizaje automático con Python: una guía para científicos de datos]. O'Reilly. https://books.google.com.pe/books?id=1-4lDQAAQBAJ

Network Development and MARC Standards Office. (2020, 8 de junio). *MARC 21 XML schema*. Library of Congress. http://www.loc.gov/standards/marcxml

Network Development and MARC Standards Office. (2021, 30 de julio). *Subject heading and term source codes*. Library of Congress. https://www.loc.gov/standards/sourcelist/subject.html

Ñaupas Paitán, H., Mejía Mejía, E., Novoa Ramírez, E., & Villagómez Paucar, A. (2014). *Metodología de la investigación cuantitativa-cualitativa y redacción de la tesis* (4.ª ed.). Ediciones de la U. https://books.google.com.pe/books?id=VzOjDwAAQBAJ

Peters, I. (2009). *Folksonomies: indexing and retrieval in Web 2.0* [Folksonomías: indexación y recuperación en la web 2.0]. De Gruyter Saur. https://books.google.com.pe/books?id=HBYg36gbnegC

Policy and Standards Division. (2013). *Library of Congress Subject Headings* [Encabezamientos de materia de la Library of Congress] (35.ª ed., Vol. 1). Library of Congress. https://books.google.com.pe/books?id=ZzKhAOH45voC

Porter, J. (2011). Folksonomies in the library: their impact on user experience, and their implications for the work of librarians. *The Australian Library Journal*, *60*(3), 248-255. https://doi.org/10.1080/00049670.2011.10722621

R Core Team. (2021). *R: a language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.r-project.org

Rahman, A. I. M. J. (2012). *Social tagging versus expert created subject headings* [Tesis de maestría, Oslo and Akershus University College of Applied Sciences]. http://eprints.rclis.org/25587

Samanta, K. S., & Rath, D. S. (2020). User-generated social tags versus librarian-generated subject headings: a comparative study in the domain of history. *DESIDOC Journal*

of Library & Information Technology, *40*(3), 176-184. https://doi.org/10.14429/djlit.40.03.15413

Samanta, K. S., & Rath, D. S. (2021). Measuring the applicability of user-generated social tags along with expert-generated LCSH descriptors in sociology: a heuristic study. *Annals of Library and Information Studies*, *68*(1), 28-38. http://nopr.niscair.res.in/handle/123456789/57114

Sánchez Nogales, E. (2020, 17 de junio). *ComunidadBNE, la plataforma colaborativa de la Biblioteca Nacional de España* [Diapositivas de PowerPoint]. RECERCAT. http://hdl.handle.net/2072/376485

Sarabia Alegría, J. M., & Pascual Sáez, M. (2012). *Curso básico de estadística para los grados en economía y administración y dirección de empresas*. Universidad de Cantabria. https://books.google.com.pe/books?id=xP5CD_YoHS0C

Sarkar, D. (2019). *Text analytics with Python: a practitioner's guide to natural language processing* [Análisis de texto con Python: guía de procesamiento de lenguaje natural para profesionales] (2.ª ed.). Apress. https://books.google.com.pe/books?id=arWZDwAAQBAJ

Shafique, F., Khan, M., Jabeen, F., & Sanila. (2019). Semantic richness of tag sets: analysis of machine generated and folk tag set. En R. Silhavy (Ed.), *Software engineering methods in intelligent algorithms: proceedings of 8th Computer Science On-line Conference 2019* (Vol. 1, pp. 35-47). https://books.google.com.pe/books?id=2cOWDwAAQBAJ

Silge, J., & Robinson, D. (2016). tidytext: text mining and analysis using tidy data principles in R. *Journal of Open Source Software*, *1*(3). https://doi.org/10.21105/joss.00037

Silge, J., & Robinson, D. (2021). *Text mining with R* [Minería de textos con R]. O'Reilly. https://www.tidytextmining.com

Srinivasa-Desikan, B. (2018). *Natural language processing and computational linguistics: a practical guide to text analysis with Python, Gensim, spaCy, and Keras* [Procesamiento de lenguaje natural y lingüística computacional: una guía práctica para el análisis de textos con Python, Gensim, spaCy y Keras]. Packt. https://books.google.com.pe/books?id=48RiDwAAQBAJ

Steele, G. (2012). The wisdom of the cataloguers: LCSH, indexer inconsistencies and collective intelligence. En A. Gilchrist, & J. Vernau (Eds.), *Facets of knowledge organization: proceedings of the ISKO UK Second Biennial Conference, 4th - 5th July, 2011, London* (pp. 61-68). Emerald. https://books.google.com.pe/books?id=QdmYwK2BhjwC

Stock, W. G., & Stock, M. (2013). *Handbook of information science* [Manual de ciencia de la información]. De Gruyter Saur. https://books.google.com.pe/books?id=d1PnBQAAQBAJ

Tchakounté, F., & Hayata F. (2017). Supervised learning based detection of malware on Android. En M. H. Au, & K-K. R. Choo (Eds.), *Mobile security and privacy: advances, challenges and future research directions* (pp. 102-154). Syngress. https://books.google.com.pe/books?id=iANaCgAAQBAJ

Torres-Moreno, J-M. (2014). *Automatic text summarization* [Resumen automático de textos]. ISTE; Wiley. https://books.google.com.pe/books?id=aPHsBQAAQBAJ

Vaidya, P., & Harinarayana, N. S. (2016). The comparative and analytical study of LibraryThing tags with Library of Congress Subject Headings. *Knowledge Organization*, *43*(1), 35-43. https://doi.org/10.5771/0943-7444-2016-1-35

Velasco de Diego, M., Llorens Morillo, J. B., & Moreiro González, J. A. (1999). Estado actual del proyecto GDA (Gestión Documental Automatizada): planteamiento teórico y descripción práctica. En F. J. García Marco (Coord.), *Organización del conocimiento en sistemas de información y documentación: actas del III Encuentro de ISKO-España, Getafe, 19 al 21 de noviembre de 1997* (pp. 317-326). Universidad de Zaragoza. https://dialnet.unirioja.es/servlet/articulo?codigo=2036358

Wagner, C., Singer, P., Strohmaier, M., & Huberman, B. (2014). Semantic stability and implicit consensus in social tagging streams. *IEEE Transactions on Computational Social Systems*, *1*(1), 108-120. https://api.semanticscholar.org/CorpusID:14632154

Wang, X., Zhang, Y., & Yamasaki, T. (2019). User-aware folk popularity rank: user-popularity-based tag recommendation that can enhance social popularity. En *MM '19: proceedings of the 27th ACM International Conference on Multimedia* (pp. 1970-1978). ACM. https://doi.org/10.1145/3343031.3350920

Wickham, H. (2014). Tidy data. *Journal of Statistical Software*, *59*(10). http://dx.doi.org/10.18637/jss.v059.i10

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D. A., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., ... Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43). https://doi.org/10.21105/joss.01686

Wickham, H. (2020). *ggplot2: elegant graphics for data analysis* [ggplot2: gráficos elegantes para el análisis de datos] (3.ª ed.). Springer. https://ggplot2-book.org

Wickham, H., & Seidel, D. (2020). *scales: scale functions for visualization*. The Comprehensive R Archive Network. https://cran.r-project.org/package=scales

Wikström, P. (2014). #srynotfunny: communicative functions of hashtags on Twitter. *SKY Journal of Linguistics, 27*, 127-152. http://urn.kb.se/resolve?urn=urn:nbn:se:kau:diva-34891

Wright, C., Ellis, S., Hicks, S., & Peng, R. D. (2021). *Tidyverse skills for data science in R* [Competencias en tidyverse para la ciencia de datos en R]. Leanpub. https://leanpub.com/tidyverseskillsdatascience

Young, H. (Ed.). (1988). *Glosario ALA de bibliotecología y ciencias de la información*. American Library Association. https://books.google.com.pe/books?id=0drtLloSlnAC

Young, J. (2009, 02 de julio). *Expanding the power of the Library's family of vocabularies: genre/form headings* [Video]. Library of Congress. https://www.loc.gov/item/webcast-4627

Zaccone, G., & Karim, M. R. (2018). *Deep learning with TensorFlow: explore neural networks and build intelligent systems with Python* [Aprendizaje profundo con TensorFlow: explora redes neuronales y construye sistemas inteligentes con Python] (2.ª ed.). Packt. https://books.google.com.pe/books?id=zZlUDwAAQBAJ

Zeng, M. L., Žumer, M., & Salaba, A. (Eds.) (2011). *Functional requirements for subject authority data (FRSAD): a conceptual model* [Requisitos funcionales para datos de autoridad de materia (FRSAD): un modelo conceptual]. De Gruyter Saur. https://books.google.com.pe/books?id=apNs97Q8yzIC

Zhao, J., Zhang, Q., Sun, Q., Huo, H., Xiao, Y., & Gong, M. (2021). FolkRank++: an optimization of FolkRank tag recommendation algorithm integrating user and item information. *KSII Transactions on Internet and Information Systems*, *15*(1), 1-19. https://doi.org/10.3837/tiis.2021.01.001

Zubiaga, A., Körner, C., & Strohmaier, M. (2011). Tags vs shelves: from social tagging to social classification. En *HT '11: proceedings of the 22nd ACM Conference on Hypertext and Hypermedia* (pp. 93-102). ACM. http://www.markusstrohmaier.info/documents/2011_HT11_Tags_vs_Shelves.pdf

Zwaard, K., Aydelott, M., Brunton, D., Crawford, M., Grotke, A., Marcou, N., Mears, J., Nagel, J., Potter, A., Rago, M., Short, S., & Sweeney, J. M. (2018, 24-30 de agosto). *Institution as social media collector: lessons learned from the Library of Congress* [Ponencia]. IFLA WLIC 2018: Transform Libraries, Transform Societies, Kuala Lumpur, Malasia. http://library.ifla.org/id/eprint/2428