La minería de textos como subsidio para la organización de la información: un estudio exploratorio

Text mining as a subsidy for information organization: an exploratory study

Mineração de texto como subsídio para organização da informação: um estudo exploratório

Rubén Urbizagastegui-Alvarado
ORCID: http://orcid.org/0000-0001-5014-801X

Bibliotecário da Universidade da Califórnia em Riverside (UCR), Estados Unidos da América.
Email: ruben@ucr.edu

RESUMEN: Se presentan seis exploraciones utilizando el paquete R disponible para la minería de textos. Estos paquetes de minería de textos se pueden utilizar para ofrecer subsidios en la construcción de encabezamientos de materias, palabras clave y/o términos de indexación para artículos de revistas. Con los paquetes textrank, slowraker y rapidraker, la coincidencia entre las palabras clave ofrecidas por la autora del documento usado como prueba alcanzó el 50%, pero al mismo tiempo los paquetes ofrecieron palabras clave complementarias como subsidios relevantes para enriquecer la terminología orientada a la recuperación de la información. Con los paquetes tm y udpipe la coincidencia entre las palabras clave ofrecidas por la autora del documento usado como prueba alcanzó el 75%. Del mismo modo, ambos paquetes ofrecieron otras palabras clave perfectamente pertinentes para enriquecer la terminología orientada a la recuperación de la información. El único paquete inadecuado fue el RKEA.

PALABRAS-CLAVE: Minería de textos; Encabezamientos de materias; Términos de indexación; textrank; slowraker; rapidraker; tm; udpipe; RKEA.

RESUMO: Seis explorações são apresentadas usando o pacote R disponível para mineração de texto. Estes pacotes de mineração de texto podem ser usados para oferecer subsídios na construção de cabeçalhos de assunto, palavras-chave e/ou termos de indexação para artigos de periódicos. Com os pacotes textrank, slowraker e rapidraker, a coincidência entre as palavras-chave oferecidas pela autora do documento utilizado como evidência chegou a 50%, mas ao mesmo tempo os pacotes ofereciam palavras-chave complementares como subsídios pertinentes para enriquecer a terminologia voltada para a recuperação em formação. Com os pacotes tm e udpipe a coincidência entre as palavras-chave oferecidas pelo autor do documento utilizado como evidência chegou a 75%. Da mesma forma, ambos os pacotes ofereciam outras palavras-chave perfeitamente pertinentes para enriquecer a terminologia voltada para a recuperação da informação. O único pacote inadequado foi o RKEA.

PALAVRAS-CHAVE: Mineração de textos; Cabeçalhos de assuntos; Termos de indexação; textrank; slowraker; rapidraker; tm; udpipe; RKEA.

ABSTRACT: Six exploits are presented using the R package available for text mining. These text mining packages can be used to provide subsidies in the construction of subject headers, keywords and/or indexing terms for journal articles. With textrank, slowraker and rapidraker packages, the coincidence between the keywords offered by the author of the document used as evidence reached 50%, but at the same time the packages offered complementary keywords as relevant subsidies to enrich the terminology focused on information retrieval. With the tm and udpipe packages, the coincidence between the keywords offered by the author of the document used as evidence reached 75%; likewise, both packages offered other perfectly pertinent keywords to enrich the terminology focused on information retrieval. The only inappropriate package was the RKEA.

KEYWORDS: Text mining; Subject headers; Indexing terms; textrank; slowraker; rapidraker; tm; udpipe; RKEA.

1 Introducción

Los encabezamientos de materias (EM) están compuestos por un grupo de palabras que capturan la esencia de los temas contenidos por un objeto informativo poseído por las bibliotecas; este objeto informativo puede ser un libro, una tesis, una publicación seriada, una película, mapas o materiales cartográficos de diversos tipos, manuscritos, archivos de computadora, recursos electrónicos, o cualquier otro material mantenido en un centro de información. Estos encabezamientos de materias se seleccionan de una lista de "autoridades" que contienen los términos de acceso a los asuntos; una especie de vocabulario controlado que se asigna como entrada adicional en los registros bibliográficos. Opera como un punto de acceso al objeto informativo y permite que el ítem al ser procurado a través del catálogo de la biblioteca o centro de información sea recuperado y posiblemente leído. Es decir, opera como un mediador entre el usuario y la información. Las decisiones tomadas sobre la representación temática de los textos pueden tener un impacto significativo en la calidad de la recuperación de la información y en la confianza de los usuarios en sus sistemas de información.

La construcción de encabezamientos de materias como puntos de acceso al objeto informativo es un proceso bastante complejo, pues en su construcción se tiene que identificar términos sinónimos y seleccionar los términos preferidos por los productores de las ciencias para usarlos como puntos de acceso. Es decir, estos encabezamientos de materias deben representar el lenguaje de sus creadores. Por ejemplo, en el caso de homónimos, identifica los múltiples conceptos expresados por esas palabras o frases. La idea central es evitar que diferentes organizadores de la información usen diferentes términos para un mismo concepto y así evitar también que el usuario utilice diferentes términos para representar el mismo tema al momento de buscar información en un centro de información o biblioteca. Allí

reside la justificación del por qué se usan las referencias cruzadas para dirigir al usuario de los términos que no se usan como EM a los términos usados para ese fin, así como a términos relacionados con el término seleccionado para representar y recuperar un asunto determinado. En resumen, el vocabulario controlado ayuda a superar los problemas inherentes al uso del lenguaje natural como puntos de acceso a los materiales informativos. Este proceso de selección, creación de EM es lento y costoso, por lo que muchas bibliotecas tienden a adoptar lenguajes documentarios elaborados en otras latitudes creando de esa manera una especie de alejamiento de la terminología usada por sus productores nativos nacionales (URBIZAGÁSTEGUI, 1994, 2022). La falta de actualización y ajuste de una autoridad temática puede ser un problema de conocimientos subyacentes; por ejemplo,

> el paradigma funcionalista de las LCSH dificulta la descripción de diferentes posturas ideológicas como la tradición dialéctica en la que ha escrito Pierre Bourdieu" (OLSON; SCHLEGL, 2021, p. 68).

Para el caso peruano,

> Son muchos los dolores de cabeza de los especialistas en procesos técnicos, en particular y de los bibliotecólogos, en general, al enfrentarse a la tarea cotidiana de indizar … [nos] ... referimos específicamente a [...] la desactualización de los temas y la cada vez más lejana relación con la realidad peruana (DORIVAL CÓRDOVA; ROJAS LAZARO, 2012, p. 27).

Por otro lado, desde la década de los 80 del siglo pasado se viene explorando la minería de datos como un proceso computacional para descubrir patrones en grandes conjuntos de datos que involucran técnicas de inteligencia artificial, el aprendizaje automático y la exploración en bases de datos. La minería de textos busca extraer información de un texto o conjunto de textos y transformarla en una estructura comprensible para su uso, es decir, es el proceso de descubrimiento de conocimiento en bases de datos bibliográficas. Esto es conocido como extracción de conocimiento de información estructurada por que la recuperación de la información se hace de campos que ya están organizados y estructurados de alguna

forma; por ejemplo, en la bases de datos bibliográficas ya están estructurados y organizados los campos del autor, título, palabras clave y eso ha dado lugar a lo que se conoce como ***minería de textos***. Sin embargo, existe otro tipo de textos que aún no están almacenados en las bases de datos bibliográficas y por lo tanto aún no están estructurados, son amorfos y difíciles de manejar algorítmicamente. Estos son los textos en formato de libros, tesis, artículos de revistas, monografías y similares aún no procesados como información bibliográfica.

En todos los campos que involucran grandes colecciones de datos sin procesar, la transformación de las observaciones en información utilizable es un objetivo implícito buscando como objetivo final la traducción de la información en conocimiento. La automatización del descubrimiento de conocimientos a partir de textos no estructurados podría subsidiar la construcción de encabezamientos de materias y vocabularios controlados más precisos a través del uso de la minería de los documentos y de la identificación del vocabulario usados por los propios creadores; de esa manera se recuperaría la propia terminología utilizada por los especialistas creadores del texto. Esta información no estructurada se encuentra en objetos físicos como libros, tesis académicas, revistas y similares objetos de información; también se encuentra en la forma de mensajes de texto en páginas web, wikis, blogs, libros electrónicos y muchos más. El tiempo disponible para que una persona recopile, lea, interprete e indexe un texto en lenguaje natural es limitada, por lo tanto, lo que se busca es la automatización de este proceso pero como subsidiaria para apoyar el proceso de tratamiento y organización de la información. La minería de textos por lo general trata con textos cuya función es la comunicación de opiniones o información fáctica, y la motivación para tratar de extraer información de dichos textos automáticamente es convincente, incluso si el éxito es solo parcial. La tarea principal de la extracción de palabras clave es identificar una palabra o frase que represente el contenido principal del texto, pues como expresión concisa de la idea principal de un documento, la palabra clave facilita la gestión, clasificación y recuperación de la información, al mismo tiempo que se convierte en un subsidio para

la elaboración de encabezamientos de materias y por lo tanto, la creación de autoridades temáticas. Por esas razones, el objetivo de este artículo es presentar algunas exploraciones utilizando la paquetería de R disponibles para la creación de palabras claves. El autor de este documento no conoce aplicaciones prácticas de la minería de textos orientados a ofrecer subsidios en esta dirección, pero si existe literatura que describa las ventajas y desventajas de la minería de textos en sus aplicaciones al campo de la bibliotecología y las ciencias de la información.

2 Revisión de la literatura

La mayoría de los trabajos publicados sobre el proceso de extracción de conocimientos en bases de datos bibliográficas y la forma cómo se puede utilizar la minería de textos en las bibliotecas para entender los patrones del comportamiento de los usuarios en el uso de los recursos de información en las instituciones apenas describen y grafican los posibles beneficios del uso de las técnicas de minería de textos en las bibliotecas. Los interesados en revisar estas descripciones básicas y generales de lo que sería la aplicación de la minería de textos al campo de la Bibliotecología y Ciencias de la Información (BCI) pueden revisar los documentos publicados por CANDÁS-ROMERO (2006); BOTTA-FERRET; CABRERA-GATO (2007), GORBEA-PORTAL (2013); CONTRERAS BARRERA (2014, 2016); JARAMILLO VALBUENA; CARDONA; FERNANDEZ (2015). También GÁLVEZ (2008) describe el auge y las limitaciones de las herramientas de análisis de la información en lenguaje natural, almacenada en bases de datos bibliográficas, como PubMed o MEDLINE. La única experiencia práctica de minería de textos parece haber sido hecha por URBIZAGÁSTEGUI (2021), quien tomando una muestra de 48 documentos publicados sobre el asunto Bibliometría Brasileña, desde 1973 hasta 2020, analizó las características textuales de esta literatura. Utilizó las técnicas de minería de textos, enfocándose principalmente en la identificación de la frecuencia del uso de los términos con el paquete tm de R. El algoritmo Latent Dirichlet Allocation (LDA)

agrupó los documentos de la muestra en 5 tópicos diferentes con la identificación de las palabras más recurrentes en cada asunto agrupado. Estos asuntos son mostrados con la construcción de un dendrograma y el análisis de clústeres realizado con el algoritmo correspondiente al método de Ward que identificó tres clusters homogéneos. Finalmente construyó una red de relaciones de las palabras o tokenes de los 48 documentos analizados. El propio URBIZAGÁSTEGUI (2021) partiendo del supuesto de que las formas en que se organiza la información, la documentación y la forma en que se comunican a los interesados reflejan y constituyen la forma en que estos problemas son representados socialmente, estudió la literatura publicada sobre arte rupestre peruano, hasta diciembre de 2020. Utilizó el software Iramuteq como herramienta para el análisis estadístico del corpus textual. Todas las contribuciones incluidas y organizadas como un solo corpus, fueron sometidas a 6 tipos de estudio: análisis de léxico clásico, clasificación jerárquica descendiente (CHD), análisis de especificidades de grupos, análisis factorial de correspondencias (AFC), análisis de similitudes y nubes de palabras. El análisis lexical clásico identificó 216 segmentos de textos formados por 8620 ocurrencias de unidades lingüísticas y de las cuales 1857 fueron lemas, 1776 palabras de formas activas; de este total de palabras activas, 414 tuvieron una frecuencia de 3 o mayor a tres usos en los textos; 81 palabras fueron formas suplementarias y 1043 palabras tuvieron una única aparición (56.17% de las formas) conocidas como hápax. El análisis de correspondencias identificó que los dos ejes principales de la coordenada cartesiana explican el 71.65% de la inercia en la distribución de los términos en clases. El análisis de similitud identificó cuatro clases léxicas que se mostraron en un dendrograma.

Como se puede ver por la literatura revisada, los aportes de la minería de textos al campo de la bibliotecología y ciencias de la información son realmente escasos y parece no haber sido explorado seriamente todavía.

3 Material y métodos

Como unidad de análisis experimental se tomó el resumen en inglés (abstract) de la tesis de Morais (2018), debido a que toda la paquetería de R está pensada para trabajar en inglés como primera aproximación se deseaba experimentar con el idioma inglés. Una experimentación con el idioma español y portugués está pendiente. En el proceso de análisis, el primer paso es copiar el resumen en inglés y usando Notepad guardarlo como texto plano en el formato UTF-8. Una vez guardado el texto se realiza la lectura del mismo con los paquetes correspondientes. En este experimento se usaron los siguientes paquetes: textrank, slowraker, tm, rapidraker, Rkea y Udpipe.

El paquete textrank (Wijffels, 2020) hace un resumen del texto clasificando oraciones y encontrando palabras clave. El resumen del texto es realizado calculando cómo se relacionan las oraciones entre sí. Esto se hace observando la terminología utilizada en las oraciones para luego establecer vínculos entre estas oraciones gramaticales. El paquete slowraker (Baker, 2017) es una versión lenta del algoritmo de extracción automática rápida de palabras clave (RAKE), que se usa para extraer palabras clave de documentos sin datos de entrenamiento. tm: Text Mining Package (Feinerer y Hornik, 2020) es un paquete desarrollado para aplicaciones de minería de texto con R. Ya Rapidraker: Rapid Automatic Keyword Extraction (RAKE) Algorithm (Baker, 2021) es una implementación en Java del algoritmo RAKE, que se puede usar para extraer palabras clave de documentos sin ningún proceso de entrenamiento. RKEA (Feinerer y Hornik, 2015) es una interfaz de R para KEA (Versión 5.0), un algoritmo que permite extraer frases clave de documentos de textos. Se puede utilizar para la indexación libre o para la indexación con un vocabulario controlado. Finalmente, Udpipe (Wijffels, 2021) es un paquete que realiza la tokenización, etiquetado de partes del discurso, lematización y análisis de dependencias de textos sin formato, también contiene funcionalidades para manipulaciones de datos de uso común en textos que se enriquecen con el análisis del investigador. Los resultados de la identificación de las palabras clave candidatas a la construcción de

encabezamientos de materias será descrito paquete por paquete. En otras palabras, estos paquetes ofrecen solo los subsidios necesarios para una toma de decisiones finales por los catalogadores, clasificadores, indexadores, es decir, los comprometidos con la organización de la información en un centro de documentación y/o biblioteca. Esta decisión final estará basada en las palabras clave identificadas por los paquetes de R y la forma en que expresan adecuadamente el asunto del texto bajo análisis. Cuando hablamos de subsidio entendemos como un enfoque del posicionamiento orgánico orientado a facilitar la identificación de palabras clave que se basan en las estructuras mentales de los creadores de la información.

4 Resultados

La autora del texto usado en este experimento, Morais (2018), ofrece como palabras clave los siguientes términos: ***Análisis de dominio, Construcción de lenguaje documental, Análisis de dominio instrumental y Lenguajes documentales***. Estos términos serán tomados como matrices de comparación para medir el éxito o fracaso de los paquetes probados en la construcción de palabras clave como subsidios para la elaboración de encabezamientos de materias en el texto.

a) Paquete textrank

Con la aplicación del paquete **textrank** se obtuvieron 30 palabras clave pero en la **Tabla 1** sólo se muestran las 25 palabras claves más importantes. Es evidente que las palabras clave escogidas serían: análisis de dominio, análisis de dominio instrumental, análisis de dominio descriptivo, lenguajes documentarios, análisis de contenido, lenguajes de representación, comunidad discursiva, garantía literaria y ciencia de la información. Como se puede observar en los datos, existe plena coincidencia entre las palabras clave construidas por el paquete **textrank** y las ofrecidas por la propia autora. En otras palabras, el subsidio ofrecido por el uso del paquete **textrank** puede ser considerado adecuado, pues muestra coincidencia en dos términos y sugiere términos complementarios pertinentes pero no considerados

por la autora. Para quienes no leen en inglés bastaría llevar estas palabras clave al **Google translate** y traducirlas al español o portugués.

Tabela 1 - Palabras clave construidas por el paquete textrank

Palabras clave	Palabras clave
instrumental domain analysis	termnological aspects
descriptive domain analysis	terminological method
domain analysis	delimitation stage
domain delimitation stage	terminological methodologies
certain domain area	information science
documentary languages	methodological applicability
notional domain	qualitative approach
respective domain	analytical approach
content analysis	discursive communities able
representation languages	practical approach
documentary linguistics	methodological tool
literary guarantee present	epistemological ontological
certain discursive community	

Fuente: Elaboración propia

b) Paquete slowraker

El paquete **slowraker** subsidió las palabras clave mostradas en la **Tabla 2**. Es evidente que las palabras clave escogidas serían: análisis de dominio; lenguajes documentarios; ciencia de la información. Después se podría jugar con una combinación de: construcción del dominio y delimitación de la investigación. En este caso hay coincidencia en dos palabras clave construidas por el paquete **slowraker** y las ofrecidas por la propia autora del documento usado como prueba. Aun así hay un aporte de dos nuevas terminologías pertinentes ofrecidas como subsidio por el uso del paquete **textrank**; por lo tanto, el resultado de la prueba se puede considerar como adecuado para este paquete.

Tabela 2 - Palabras clave sugeridas por el paquete slowraker

Palabras clave	Frecuencia
domain analysis	11
documentary languages	8
construction	6
domain	4
research	4
delimitation	3
information science	3

Fuente: Elaboración propia

c) Paquete tm

El paquete **tm** identificó las palabras clave mostradas en la **Tabla 3**. Es evidente que las palabras clave escogidas serían nuevamente: análisis de dominio, lenguajes documentarios, construcción documentaria, delimitación del dominio, estado de la delimitación, ciencia de la información, análisis del paradigma, construcción del dominio e investigación del dominio. En este caso nuevamente se verifica una alta coincidencia entre las palabras clave construidas por el paquete **tm** y las ofrecidas por la propia autora. Se verifica también que hay un aporte de terminologías identificadas por el paquete **tm**. Se puede concluir afirmando que el subsidio que ofrece este paquete es altamente positivo.

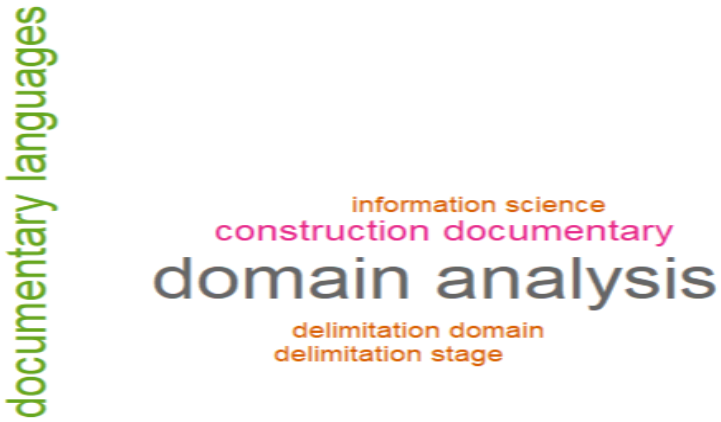
Tabla 3 - Palabras clave construidas por el paquete tm

Palabras clave	Frecuencia
domain analysis	13
documentary languages	8
construction documentary	5
delimitation domain	3
delimitation stage	3
information science	3
analysis paradigm	2
domain construction	2
domain delimitation	2
investigate domain	2

Fuente: Elaboración propia

Con este paquete se puede construir también una nube de palabras cuyo resultado se puede observar en la Figura 1. El tamaño de las palabras muestra el énfasis dado a las frases en el texto así como en la construcción de las palabras clave.

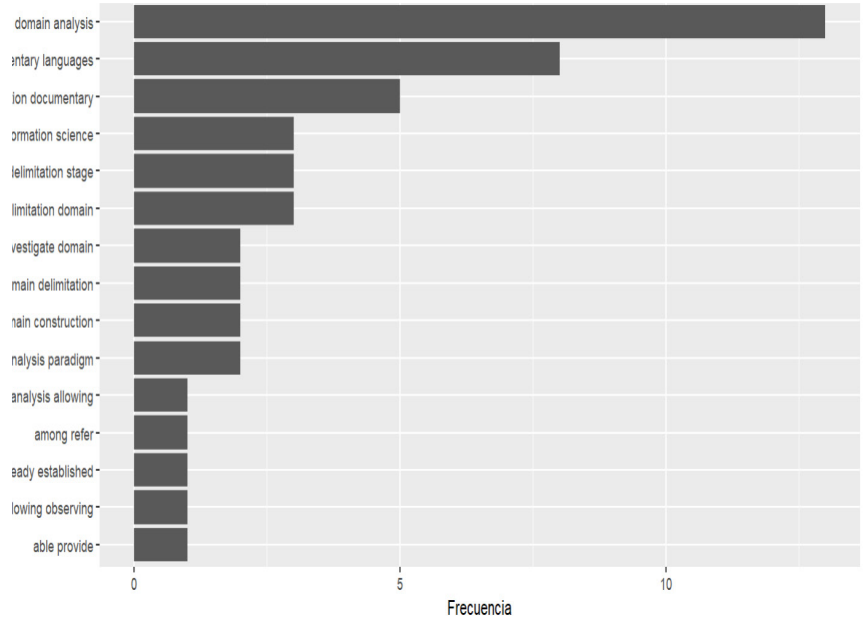
Figura 1 - Nube de palabras identificadas por el paquete tm



Fuente: Elaboración propia

Con este paquete se pueden también crear gráficos de bigramas y trigramas que pueden ayudar mucho más a la selección de los encabezamientos de materias o de las palabras clave pertinentes del documento analizado. La Figura 2 muestra los bigramas de palabras clave construidos con el paquete tm.

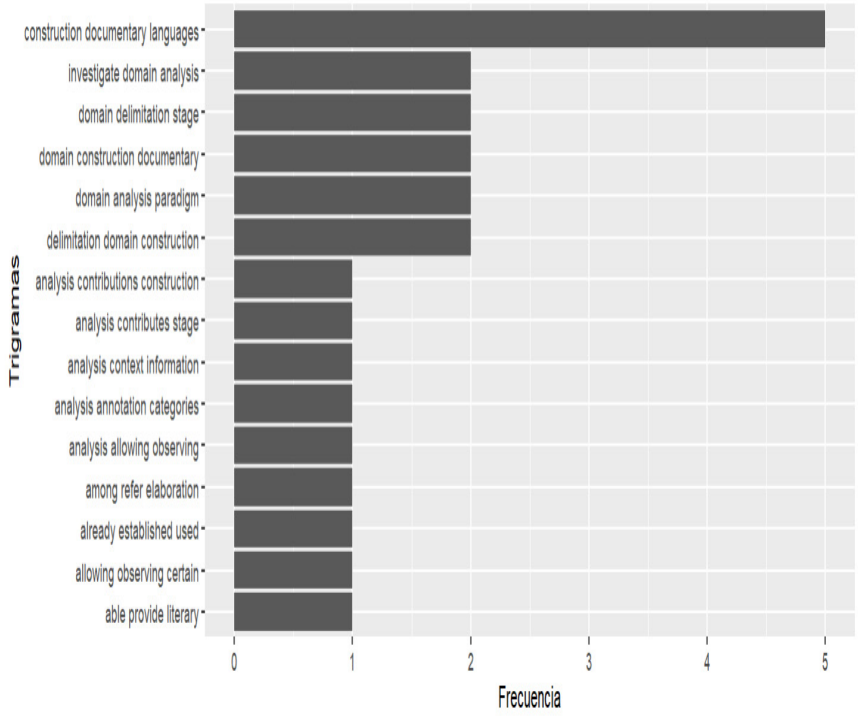
Figura 2 - Bigramas de palabras clave elaboradas por el paquete tm



Fuente: Elaboración propia

La Figura 3 muestra los trigramas de palabras clave construidos con el paquete tm. Estas palabras clave también pueden ayudar en la construcción de los encabezamientos de materias o las palabras clave de los documentos bajo análisis.

Figura 3 - Trigramas de palabras clave elaboradas por el paquete tm



Fuente: Elaboración propia

d) Paquete rapidraker

El paquete rapidraker identificó las palabras clave mostradas en la Tabla 4. Este paquete identificó 87 palabras clave pero aquí se muestran solamente aquellas con puntajes mayores a 4.0. Nuevamente, es evidente que las palabras clave escogidas serían: análisis descriptivo de dominio, etapa de delimitación del dominio, análisis instrumental del dominio, análisis de contenido, análisis de dominio, herramientas bibliométricas, sistemas de clasificación y comunidades discursivas. En este caso nuevamente se verifica una alta coincidencia entre las palabras clave construidas por el paquete rapidraker y las ofrecidas por la propia autora. Se puede concluir afirmando que el subsidio que ofrece este paquete también es altamente positivo.

Tabla 4 - Palabras clave construidas por el paquete rapidraker

Palabras clave	frecuencia	Puntaje
descriptive domain analysis	1	7.066667
domain delimitation stage	2	5.833333
instrumental domain analysis	1	5.733333
content analysis	1	4.066667
domain analysis	11	4.066667
bibliometric tools	1	4.000000
classification systems	1	4.000000
discursive communities	1	4.000000
discursive community	1	4.000000
documentary linguistics	1	4.000000
domain area	1	4.000000
initial study	1	4.000000
literary guarantee	1	4.000000
main objective	1	4.000000
notional domain	1	4.000000
notional system	1	4.000000
representation purposes	1	4.000000
respective domain	1	4.000000
scientific method	1	4.000000
specific objectives	1	4.000000
theoretical reference	1	4.000000
theoretical tendencies	1	4.000000

Fuente: Elaboración propia

e) Paquete RKEA

El paquete RKEA identificó las palabras clave mostradas en la Tabla 5. Es evidente que las palabras clave escogidas serían nuevamente: análisis de dominio, estado de la delimitación del dominio, delimitación del dominio. En este caso se verifica una baja coincidencia entre las palabras clave construidas por el paquete y las ofrecidas por la propia autora.

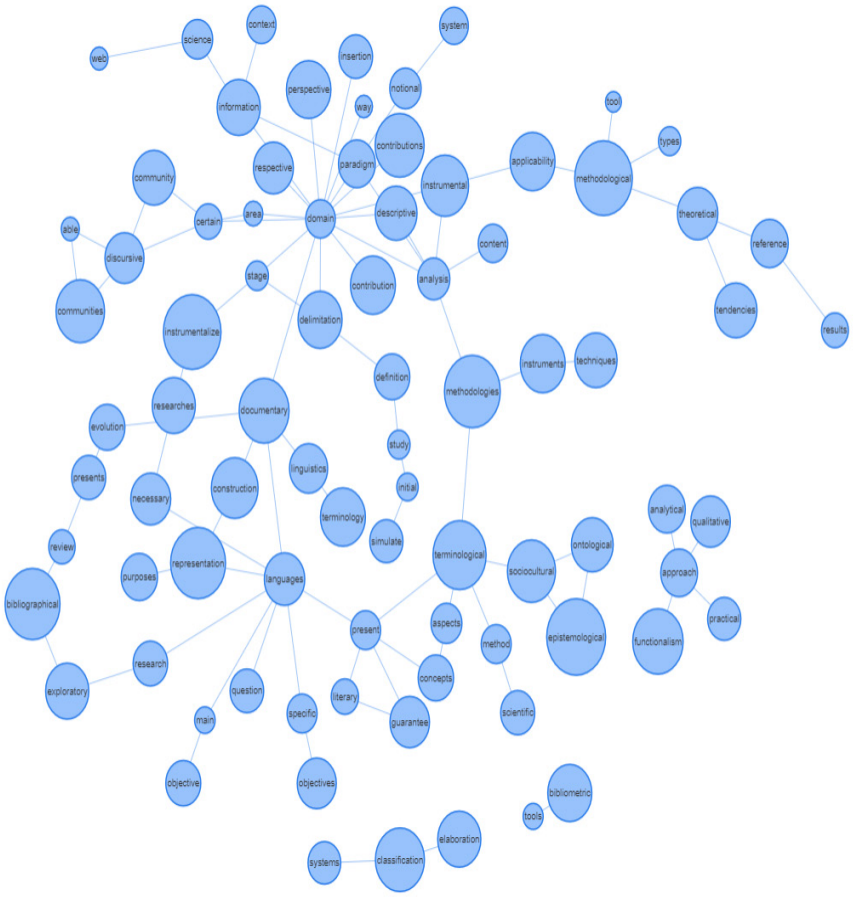
Tabla 5 - Palabras clave construidas por el paquete RKEA

Palabras clave
research
points
domain
domain analysis
analysis
analysis in the domain delimitation
domain delimitation
domain delimitation stage
delimitation
delimitation stage

Fuente: Elaboración propia

Con este paquete se puede construir también una red de relaciones de las palabras que se puede observar en la Figura 4. El tamaño de las palabras muestra el destaque dado a las frases en el texto así como en la construcción de las palabras clave.

Figura 4 - Red de palabras clave elaboradas por el paquete RKEA



Fuente: Elaboración propia

f) Paquete udpipe

El paquete udpipe identificó las palabras clave mostradas en la Tabla 6. Es evidente que las palabras clave escogidas serían nuevamente: análisis de dominio, lenguajes documentarios, construcción de lenguajes, construcción documentaria y estado de la delimitación. En este caso nuevamente se verifica una alta coincidencia entre las palabras clave construidas por el paquete udpipe y las ofrecidas por la propia autora del documento usado como prueba. Se verifica también que hay un aporte de terminologías identificadas por este paquete y se puede concluir afirmando que el subsidio que ofrece este paquete es altamente positivo.

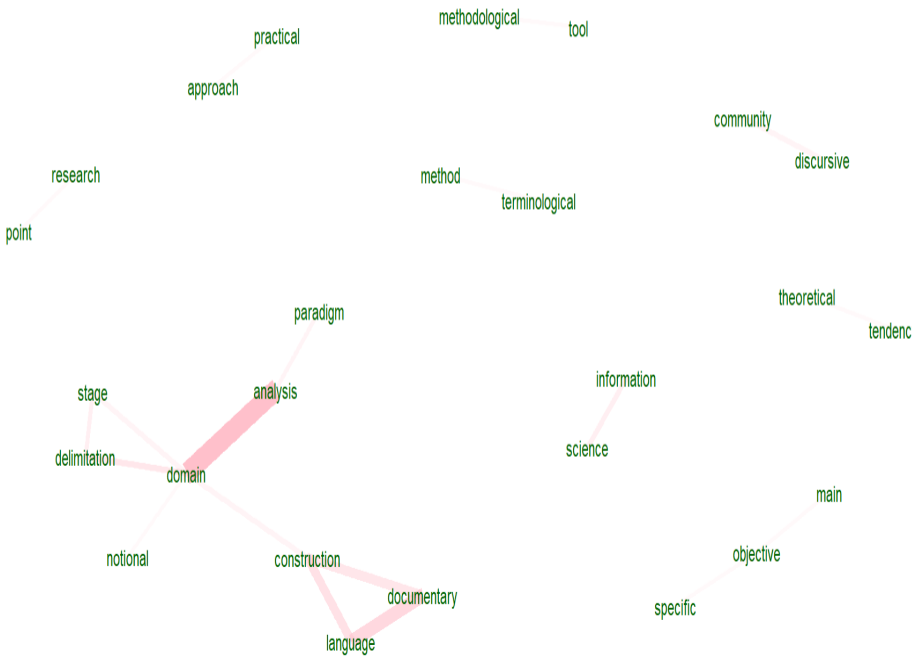
Tabla 5 - Palabras clave construidas por el paquete Udpipe

Número	Término1	Término2	Coocurrencias
1	domain	analysis	13
2	documentary	language	8
3	construction	language	6
4	construction	documentary	5
5	delimitation	stage	3
6	information	science	3

Fuente: Elaboración propia

Con este paquete se puede construir también una red de relaciones de las palabras que se puede observar en la Figura 5. El grosor de los lazos entre las palabras indica la intensidad de las coocurrencias en el texto original; véase, por ejemplo, como el centro del texto es el análisis de dominio, luego la construcción del lenguaje documentario; después delimitación del dominio, comunidad discursiva y ciencia de la información con mayor intensidad mostrada por los lazos pero que pueden ayudar a construir los encabezamientos de materias o palabras clave.

Figura 5 - Red de palabras clave elaboradas por el paquete udpipe



Fuente: Elaboración propia

Resumiendo, con la excepción de RKEA que no ofrece un subsidio adecuado para la construcción de encabezamientos de materias o palabras clave, todos los otros paquetes muestran un comportamiento adecuado en el subsidio de informaciones dirigido al objetivo perseguido. Sin embargo, los paquetes tm y udpipe parecen ofrecer los mejores resultados. Estos paquetes son subsidiados con dos palabras clave no consideradas por la autora del documento usado como prueba en esta exploración: comunidades discursivas y ciencia de la información, que complementan adecuadamente los asuntos tratados por la autora del documento usado como prueba para comparación.

5 Conclusiones

Los resúmenes de los documentos académicos son esenciales para el proceso de difusión y comunicación de la actividad científica, ya que representan, de forma condensada, el contenido temático de esos documentos. Representar adecuadamente los asuntos contenidos en esos documentos es fundamental para la recuperación de la información; los términos de indexación o palabras clave asociados a los resúmenes son los principales productos que garantizan esa recuperación. El objetivo del resumen es promover la circulación de las ideas objetivadas en los documentos a nivel nacional e internacional; esto se logra a través de su inclusión en diversos soportes informativos y bases de datos bibliográficas. Por lo tanto, si el resumen está bien elaborado no solo es una herramienta que apoya al investigador en la elección de sus fuentes, sino que es fundamental para la construcción de las palabras clave que representarán adecuadamente el asunto comunicado en el documento. Estas palabras claves pueden servir también como subsidios para la elaboración de tesauros y vocabularios controlados con la minería de textos como soporte para la actualización de estos instrumentos. Finalmente, la introducción del uso de la minería de textos en el campo de la ciencia de la información y bibliotecología obliga a repensar el rol del profesional dedicado a la organización de la información y del conocimiento. Su rol en la construcción de la representación documental no puede seguir siendo el "copiador" de encabezamientos de materias, descriptores y palabras clave construidos en otras realidades históricas y para otros contextos sociales para pasar ser el planificador y arquitecto constructor de los vocabularios controlados propios de su entorno histórico local, regional o nacional y dejar de ser apenas el reproductor de encabezamientos o vocabularios elaborados en otras latitudes, para otras realidades sociales y el difusor de esas influencias funcionalistas como hasta ahora.

La minería de textos como un área de estudio del procesamiento de los do-

cumentos textuales, se encarga de desarrollar y utilizar métodos para facilitar la extracción de conocimientos útiles para la organización de la información. En este trabajo se probaron seis paquetes disponibles en el Proyecto R y los resultados parecen prometedores. Con los paquetes textrank, slowraker, rapidraker las coincidencias entre las palabras clave ofrecidas por la autora del documento usado como prueba alcanzó el 50%, pero al mismo tiempo los paquetes ofrecieron palabras clave complementarias como subsidios pertinentes para enriquecer la terminología dirigida a la recuperación de la información. Con los paquetes tm y udpipe la coincidencia entre las palabras clave ofrecidas por la autora del documento usado como prueba alcanzó el 75%. Del mismo modo, ambos paquetes ofrecieron otras palabras clave perfectamente pertinentes para enriquecer la terminología dirigida a la recuperación de la información. El único paquete con resultados no adecuados para esta tarea fue el RKEA. Sin embargo, mayores pruebas y experimentos son necesarios en esta dirección para validar y generalizar los resultados obtenidos. Una exploración usando documentos escritos en el idioma español y portugués está en proceso y sus resultados serán comunicados oportunamente.

Referencias:

BAKER, Christopher. slowraker: A Slow Version of the Rapid Automatic Keyword Extraction (RAKE) Algorithm, 2017. R package version 0.1.1. Disponível em: https://CRAN.R-project.org/package=slowraker. Acesso em: Acesso em: 03 maio 2022.

BAKER, Christopher. rapidraker: Rapid Automatic Keyword Extraction (RAKE) Algorithm, 2021. R package version 0.1.3. Disponível em: https://CRAN.R-project.org/package=rapidraker. Acesso em: 04 maio 2022.

BOTTA-FERRET, Eleazar; CABRERA-GATO, Jania E. Minería de textos: una herramienta útil para mejorar la gestión del bibliotecario en el entorno digital. ACIMED, Ciudad de La Habana, v. 16, n. 4, oct. 2007. Disponível em: http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S1024-94352007001000005. Acesso em: 03 maio 2022.

CANDÁS-ROMERO, Jorge. "Minería de datos en bibliotecas: bibliominería. BiD: textos universitaris de biblioteconomia i documentació, v. 17, 2006. https://bid.ub.edu/sites/bid9/files/pdf/17canda1.pdf

CONTRERAS BARRERA, Marcial. Minería de texto: una visión actual. Biblioteca Universitária, v. 17, no. 2, p. 129-138, 2014. https://brapci.inf.br/index.php/res/v/51778 . Acesso em: 03 jun. 2022.

CONTRERAS BARRERA, Marcial. Minería de texto en la clasificación de material bibliográfico. Biblios, n. 64, p. 33-43, 2016. Disponível em: https://www.redalyc.org/journal/161/16148511003/html/. Acesso em: 03

abr. 2022.

DORIVAL CÓRDOVA, Rosa; ROJAS LAZARO, Carlos Javier. El uso de los sistemas tradicionales de organización del conocimiento en las bibliotecas peruanas. Biblios: Journal of Librarianship and Information Science, n. 46, p. 26-32, 2012. Disponível em: http://biblios.pitt.edu/ojs/index.php/biblios/article/download/38/93 . Acesso em: 03 abr. 2022.

FEINERER, Ingo; HORNIK, Kurt. tm: Text Mining Package. 2020. R package version 0.7-8. Disponível em: https://CRAN.R-project.org/package=tm. Acesso em: 03 abr. 2022.

FEINERER, Ingo; HORNIK, Kurt. RKEA: R/KEA Interface, 2015. R package version 0.0-6. Disponível em: https://CRAN.R-project.org/package=RKEA. Acesso em: Acesso em: 03 abr. 2022.

GÁLVEZ, Carmen. Minería de textos: la nueva generación de análisis de literatura científica en biología molecular y genómica. Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação, v. 13, no. 25, 1-14, 2008. Disponível em: https://www.redalyc.org/pdf/147/14702502.pdf. Acesso em: 03 mar. 2022.

GORBEA-PORTAL, Salvador. Tendencias transdisciplinarias en los estudios métricos de la información y su relación con la gestión de la información y del conocimiento. Perspectivas em Gestão & Conhecimento, v. 3, no. 1, p. 13-27, 2013. Disponível em: https://brapci.inf.br/index.php/res/v/53087. Acesso em: 13 mar. 2022.

JARAMILLO VALBUENA, Sonia; CARDONA, Sergio Augusto; FERNANDEZ, Alejandro. Minería de datos sobre streams de redes sociales, una herramienta al servicio de la Bibliotecología. Información, cultura y sociedad, n. 33, p. 63-74, 2015. Disponível em: http://www.scielo.org.ar/scielo.php?script=sci_arttext&pid=S1851-17402015000200005. Acesso em: 05 mar. 2022.

MORAIS, Natanna Santana de. A análise de domínio na construção de linguagens documentárias. Informação em Pauta, Fortaleza, CE, v. 3, n. 2, p. 140-141, jul./dez. 2018. Disponível em: http://www.periodicos.ufc.br/informacaoempauta/article/view/39572. Acesso em: 05 mar. 2022.

OLSON, Hope A.; SCHLEGL, Rose. Standardization, objectivity, and user focus: A meta-analysis of subject access critiques. Cataloging & classification quarterly, v. 32, no. 2, p. 61-80, 2001. Disponível em: https://www.tandfonline.com/doi/abs/10.1300/J104v32n02_06. Acesso em: 03 maio 2022.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. R version 4.1.2. 2021. Dispoível em: https://www.R-project.org/. Acesso em: 03 maio 2022.

URBIZAGÁSTEGUI-ALVARADO, Rubén. La bibliometría brasileña: minería de textos. Revista ACB: Biblioteconomía em Santa Catarina, v. 26, no. 1, p. 1-18, 2021. Disponível em : https://revista.acbsc.org.br/racb/article/view/1768. Acesso em: 07 mar. 2022.

URBIZAGÁSTEGUI-ALVARADO, Rubén. Arte rupestre peruano: análisis textométrico. En prensa.

URBIZAGÁSTEGUI-ALVARADO, Rubén. Cataloging Pierre Bourdieu's books. Cataloging & Classification Quarterly, v. 19, no. 1: 89-105, 1994. Disponível em:
https://www.tandfonline.com/doi/abs/10.1300/J104v19n01_07. Acesso em: 06 mar. 2022.

URBIZAGÁSTEGUI-ALVARADO, Rubén. Encabezamientos de materias: develando la organización del conocimiento. Revista Prefacio (Córdoba, Argentina), v. 5, n. 8, p. 79-98, 2022. Disponível em:
https://revistas.unc.edu.ar/index.php/PREFACIO/issue/view/2491. Acesso em: 05 mar. 2022.

WIJFFELS, Jan. textrank: Summarize Text by Ranking Sentences and Finding Keywords, 2020. R package version 0.3.1. Disponível em: https://CRAN.R-project.org/package=textrank. Acesso em: 08 maio 2022.

WIJFFELS, Jan. udpipe: Tokenization, Parts of Speech Tagging, Lemmatization and Dependency Parsing with the 'UDPipe' 'NLP' Toolkit, 2021. R package version 0.8.8. Disponível em: https://CRAN.R-project.org/package=udpipe. Acesso em: 05 mar. 2022.