Revista ACB: Biblioteconomia em Santa Catarina, Florianópolis, v. 26, n. 1, p. 1-18, jan./ abr., 2021.

LA BIBLIOMETRÍA BRASILEÑA: MINERÍA DE TEXTOS

Rubén Urbizagástegui-Alvarado[1]

RESUMEN

Tomando una muestra de 48 documentos publicados sobre el asunto Bibliometría Brasileña, desde 1973 hasta 2020, se analiza las características textuales de esta literatura publicada. Se utiliza 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 pertinente y adecuado a los documentos. El análisis de clústeres fue realizado con el algoritmo correspondiente al método de Ward que identificó tres clústeres homogéneos. Finalmente se construyó una red de relaciones de las palabras o tokenes de los 48 documentos analizados es este trabajo.

Keywords: Bibliometría. Informetría. Cienciometría. Brasil. Minería de textos

RESUMO

Tomando uma amostra de 48 documentos publicados sobre a Bibliometria Brasileira, de 1973 a 2020, são analisadas as características textuais dessa literatura publicada. Foram utilizadas técnicas de mineração de textos, com foco principalmente na identificação da frequência de uso dos termos com o pacote tm de R. O algoritmo Latent Dirichlet Allocation (LDA) agrupou os documentos da amostra em 5 tópicos diferentes com a identificação das palavras mais recorrentes em cada assunto agrupado. Essas questões são mostradas com a construção de um dendrograma relevante e adequado para os documentos. A análise de clusteres com o algoritmo correspondente ao método de Ward identificou três clusteres homogêneos. Finalmente, foi construída uma rede de relações das palavras ou tokens dos 48 documentos analisados neste trabalho.

Keywords: Bibliometría. Informetría. Cienciometría. Brasil. Minería de textos

ABSTRACT

Taking a sample of 48 documents published on the Brazilian Bibliometry issue, from 1973 to 2020, the textual characteristics of this published literature are analyzed. Text mining techniques are used, focusing mainly on the identification of the frequency of use of the terms with R's tm package. The Latent Dirichlet Allocation (LDA) algorithm grouped the sample documents into 5 different topics with the identification of the most recurring words in each grouped subject. These issues are shown with the construction of a relevant and appropriate dendrogram for the documents. The cluster analysis with the algorithm corresponding to Ward's method identified three homogeneous clusters. Finally, a network of relationships of the words or tokens of the 48 documents analyzed in this work was built.

[1] Doctor em Ciencia de la Información. Universidad de California en Riverside. Riverside, CA 92521 – 5900
https://orcid.org/0000-0001-5014-801X - ruben@ucr.edu
Revista ACB: Biblioteconomia em Santa Catarina, Florianópolis (Brasil) - ISSN 1414-0594
Esta obra está licenciada sob uma Licença Creative Commons.

Keywords: Bibliometrics. Informetrics. Scientometrics. Brazil. Text mining

1 INTRODUCCIÓN

La minería de datos es un proceso de descubrimiento automatizado de información útil ocultos en grandes volúmenes de datos (big data). Implica la aplicación y uso de ciertas técnicas y herramientas para transformar de forma automatizada los datos en informaciones beneficiosas. Estas técnicas y herramientas son los asuntos en los que se enfoca el llamado descubrimiento de conocimiento en bancos de datos (Knowledge Discovery in Databases – siglas en inglés KDD). La minería de datos es una etapa del KDD que requiere de un proceso interactivo en la que el usuario interviene y controla el curso de las actividades, pero también requiere de un proceso iterativo por ser una secuencia finita de operaciones en las que cada una de estas operaciones ejecutadas son independientes de los resultados de aquellas ejecutadas anteriormente. La minería de textos es un paradigma de la programación de datos con capacidad para entender el lenguaje natural de los textos con el fin de encontrar sus imprecisiones e incertezas y, para conseguir transformar el texto en algo que sea entendible por una computadora, envuelve varias áreas de la informática, como la minería de los datos, el aprendizaje de la máquina, la recuperación de la información, el lenguaje computacional y la estadística (MACHADO et al., 2010). La dificultad que presenta se encuentra en la percepción e interpretación apropiada de los diversos factores que se observan a lo largo del proceso de la minería de textos, además de la dificultad de integrar las interpretaciones que apoyen la toma de decisiones en relación con cada contexto y dejando a cargo del "analista humano" la responsabilidad de orientar y ejecutar los procesos de este conocimiento a ser repasado para la gestión estratégica (GOLDSCHIMDT; PASOS, 2005). Existen, pues, muchas informaciones que requieren un mayor grado de entendimiento que una simple evaluación, ya que muchos sistemas complejos no pueden ser comprendidos sin un análisis crítico de las relaciones subyacentes a esas informaciones y sus relaciones.

Naturalmente, una interpretación crítica de los textos es una tarea difícil, pero con el desarrollo de las bibliotecas digitales y el internet, los investigadores pueden acceder y recuperar fácilmente grandes cantidades de textos, imágenes y materiales multimedia en línea para sus investigaciones. Esos archivos proporcionan la materia prima, pero los investigadores aún necesitan confiar en sus propias notas, documentos y sus memorias para encontrar hechos "interesantes" que respalden o contradigan las hipótesis que se plantearon; por esa razón es que, en el campo de las ciencias y humanidades, las computadoras se usan esencialmente para acceder a documentos de texto, pero raramente para apoyar su interpretación y la búsqueda de nuevas hipótesis de trabajo. Mas recientemente se está buscando integrar un conjunto de funcionalidades de la minería de textos y visualización como una mecanismo que estimule nuevos conocimientos y descubrimientos; para eso se combina una visión general de la colección de textos completos, la ordenación de patrones recurrentes por frecuencias o longitudes, la búsqueda de patrones de palabras múltiples con vacíos estructurales, la comparación y la contrastación de las características de diferentes patrones de textos para mostrar patrones en el contexto donde aparecen, buscando su distribución en diferentes niveles de granularidad, es decir, a través de colecciones de textos o documentos (DON et al., 2007). Debido a que la mayor parte de la información científica se encuentra en la literatura publicada e indexada en las bases de datos bibliográficas, es necesario la aplicación de nuevos métodos de procesamiento, acceso y recuperación de la información. La minería de textos surge como una herramienta que sirve de soporte para el descubrimiento de los conocimientos ocultos en los datos almacenados en las bases de datos bibliográficas, por eso se define como el descubrimiento de

Esta obra está licenciada sob uma Licença Creative Commons.

conocimientos, a partir de patrones observables de datos estructurados, en bases de datos relacionales, comúnmente denominado como *Knowledge Discovery in Databases* (KDD). La minería de textos se orienta a la extracción del conocimiento a partir de datos no-estructurados en lenguaje natural almacenados en las bases de datos textuales.

El volumen y la diversidad de la literatura científica crece todos los días y se manifiesta en la abundante literatura científica indexada en diferentes bases de datos bibliográficas, repositorios locales y regionales. La minería de datos intenta descubrir información oculta en esa literatura científica, a la que no se puede acceder mediante simples técnicas estadísticas. Las técnicas de minería de textos representan un subconjunto importante de la minería de datos que tiene como objetivo extraer el conocimiento de datos textuales no estructurados o semiestructurados. La combinación de técnicas de minería de textos y el análisis bibliométrico puede ser explorada para descubrir patrones más invisibles en los campos de investigación que se requieran. La aplicación de las técnicas de minería de textos al campo de la bibliometría brasileña constituye uno de los desafíos en esta prometedora área de investigación de análisis de textos.

La minería de textos ha sido escogida por el mundo académico como una línea útil de investigación y cada vez más existe mayor un interés de la comunidad académica en el procesamiento automatizado del lenguaje natural. Este creciente interés es motivado por la necesidad generalizada de aplicaciones relacionadas con la recuperación de la información, la extracción de datos, ~~el~~ resumir documentos, descubrir patrones, asociaciones, reglas y realizar análisis cualitativos y cuantitativos en documentos de texto. A partir de esta perspectiva, surgieron diversos frentes de investigaciones con el perfeccionamiento de técnicas dirigidas a la selección de características, para la representación del conocimiento y la identificación de diferentes tópicos corrientes en la literatura dispersa en repositorios internacionales, regionales y nacionales. Aprovechando el auge que está experimentando la minería de textos como instrumento para el descubrimiento del significado que poseen la gran cantidad de datos bibliográficos almacenados en las bases de datos bibliográficas o textuales, el objetivo de este artículo es analizar la literatura publicada sobre las metrías (bibliometría, informetría, cienciometría, patentometría, archivometría, etc.) en el Brasil. El periodo escogido se extiende desde los primeros trabajos publicados en 1973 hasta diciembre de 2020, un periodo extenso como para esperar que la literatura publicada se acumule y sedimente.

No está demás insistir en que la posibilidad de enriquecer las exploraciones bibliométricas mediante el procesamiento del texto de los artículos

> … ofrece un nuevo campo de investigación, donde surgen los principales problemas en torno a la organización y estructura del texto, la extracción de información y su representación a nivel de metadatos. Recientemente, la disponibilidad cada vez mayor de conjuntos de datos y artículos en texto completo y formatos legibles por máquina ha hecho posible un cambio de perspectiva en el campo de la bibliometría. Desde bases de datos de preimpresos hasta los movimientos Open Access y Open Science, el desarrollo de plataformas online como ArXiv, CiteSeer o PLoS, etc., contribuyen en gran medida a facilitar la experimentación con conjuntos de datos de artículos, posibilitando la realización de estudios bibliométricos no solo considerando los metadatos de los artículos, sino también su contenido de texto completo (ATANASSOVA; BERTIN; MAYR, 2019, p. 1).

Por consiguiente, este trabajo busca contribuir al desarrollo de estudios sobre la clasificación de la información textual aplicada al portugués brasileño, ya que la cantidad de información registrada y

Revista ACB: Biblioteconomia em Santa Catarina, Florianópolis (Brasil) - ISSN 1414-0594

Esta obra está licenciada sob uma Licença Creative Commons.

producida en este idioma crece continuamente y es necesario explorar los medios específicos para su organización y consecuente recuperación en los medios digitales. Cada uno de los pasos envueltos en este proceso usa técnicas matemáticas capaces de ejecutar y realizar recomendaciones factibles. Se utiliza el lenguaje de programación R, dado que, a pesar de ser un lenguaje de programación de propósito general, cuenta con un ecosistema significativo para la minería de textos y procesamiento de lenguaje natural, lo cual facilita la implementación de las técnicas sugeridas en los pasos de la propuesta metodológica. Es importante destacar que el uso de técnicas matemáticas y del lenguaje de programación, permite un análisis más claro y libre de valores de los investigadores involucrados. Por lo tanto, aprovechando la construcción de la base de datos mencionada anteriormente es oportuno analizar los diferentes aspectos que presentan estos textos, especialmente el estudio de las características textuales de la literatura publicada focalizándose en las siguientes cuestiones:

a) ¿Con la técnica de minería de textos, es posible identificar las palabras más frecuentes que representan los asuntos contenidos en los documentos publicados sobre bibliometría brasileña?

b) ¿El algoritmo Latent Dirichlet Allocation (LDA) de R puede identificar y clasificar automáticamente estos documentos en grupos asuntos similares?

c) ¿El análisis de clústeres como método estadístico multivariante de clasificación automática de datos creará grupos de documentos homogéneos a partir de documentos quizás heterogéneos?

Para lograr el objetivo propuesto, este artículo está organizado en seis partes. En la primera se presenta una introducción al tema, se detalla el problema y se formulan las preguntas de investigación. En la segunda se revisa la literatura publicada sobre este asunto. En la tercera se describe la metodología empleada con énfasis en la forma de recolección de los datos y la forma de medición de estos. En la cuarta parte se exponen los resultados. En la quinta se proponen las conclusiones y discusiones de los resultados obtenidos. Por último, se presenta la bibliografía revisada para la elaboración de este trabajo.

2 MARCO TEORICO Y REVISIÓN DE LA LITERATURA

Una introducción didáctica a la minería de textos, casos en los que se han usado y los resultados obtenidos puede ser consultado en Aranha y Passos (2006). También Alazmi y Alazmi (2012) ilustran cómo la minería de datos y las técnicas de visualización se utilizan en grandes bases de datos para encontrar patrones y características ocultas. Examinan las tendencias actuales, las herramientas principales y la aplicación de tales tecnologías, y sus posibles usos futuros, luego revisan las tareas y técnicas utilizadas y disponibles en el mercado para llevar a cabo este proceso. Para Pezzini (2017) la minería de textos tiene potencial para ser muy bien explotada comercialmente, no sólo por la gran variedad de informaciones comerciales que se almacenan en forma de texto, sino también debido a su capacidad de ser aplicada en diferentes áreas del conocimiento. La minería de textos puede proporcionar automatización para varios servicios que son hechos por seres humanos, consecuentemente, la disminución de los costos y la rapidez en estos procesos. Ejemplos de esto es el análisis automático de sentimientos en encuestas de opinión pública, las cuales cuando son realizadas por personas tardan mucho más tiempo que si se realizan con un algoritmo de minería de textos. Luiz y Martins (2015) en un primer momento, realizan un levantamiento teórico acerca del descubrimiento de conocimientos, sus características y las teorías existentes en la literatura especializada, además de un seguimiento sobre la

Revista ACB: Biblioteconomia em Santa Catarina, Florianópolis (Brasil) - ISSN 1414-0594

Esta obra está licenciada sob uma Licença Creative Commons.

obtención, tratamiento, disponibilidad, integridad y seguridad de la información para las empresas. En un segundo momento, realizan una especie de estudio de caso, donde ofrecen ejemplos prácticos de aplicación del KDD y los resultados obtenidos con esta aplicación.

La mayoría de las investigaciones empíricas de análisis de textos han sido realizados en campos diferentes a la ciencia de la información. Por ejemplo, Junior y Gomes (2012) utilizaron el software ***Rapidminer*** y analizaron 4336 artículos en idioma portugués publicados en dos congresos realizados en el año 2011 en ingeniería de la producción. Presentan tres métodos tradicionales de aprendizaje de máquina (Naïve Bayes, k-NN y SVM) y proponen un método de agrupamientos para realizar la categorización de artículos del área de ingeniería de la producción. Concluyen que el método de agrupamientos propuesto obtuvo un mejor desempeño en las métricas de exactitud, precisión y alcance, que los métodos tradicionales. Corrêa et. al. (2012) describen del uso de la minería de textos como una búsqueda exploratoria en la fase de levantamiento bibliográfico en el área de medicina, más precisamente sobre el cáncer de cabeza y cuello. Adoptaron el método no supervisado para la extracción de jerarquías de los asuntos útiles para organizar y clasificar textos de un dominio específico. Para sobrellevar el aprendizaje no supervisado de jerarquías de asuntos a partir de los textos, utilizaron la herramienta TORCH (Topic Hierarchies). Los experimentos realizados a partir de artículos reales provenientes de PubMED, demostraron que la aplicabilidad del método no supervisado es eficaz para extraer tópicos de niveles más genéricos. Sin embargo, para identificar temas más específicos e innovadores para los especialistas del área es necesario realizar un preprocesamiento de los textos más apropiado, contando incluso con la interacción de un especialista humano.

Amadio y Procaccino (2016) conscientes de que sitios de reseñas en línea, como ***TripAdvisor***, brindan a los consumidores un poder sin precedentes para encontrar productos y servicios que satisfagan sus necesidades específicas, además de brindar retroalimentación a los proveedores, realizaron un análisis del tipo FODA (del inglés SWOT, strengths, weaknesses, opportunities, and threats) sobre las fortalezas, debilidades, oportunidades y amenazas, y segmentación de clientes en la industria hotelera. Este tipo de análisis generalmente se analiza en términos de variables como los beneficios buscados, la preferencia por productos con ciertas características, el uso del producto y la sensibilidad con relación al precio; por eso buscaban una comprensión exhaustiva de los gustos de los clientes y su comportamiento de compra. Esa comprensión ayuda a las empresas a llegar a varios grupos, usando las diferencias para guiar diferentes estrategias para aumentar la participación de las estrategias de ofertas dirigidas a cada grupo. Los autores utilizaron la base de datos ***ReviewMap*** para desarrollar una lista de las características ofrecidas por las diferentes empresas en el mercado y utilizaron la técnica probabilística de modelado de asuntos Latent Dirichlet Allocation (LDA) recopilaron datos sobre la forma en que los clientes perciben aquellas características ofrecidas, ya que es muy común que los proveedores ajusten sus ofertas a las revisiones individuales en un esfuerzo por reforzar las experiencias positivas o enmendar las negativas. Braga (2016) presenta un enfoque para la generación semiautomática de una taxonomía a partir de la coocurrencia de conceptos. Asegura que además de ser un paso en el proceso de creación de ontologías, esta técnica puede ser útil para una mejor comprensión del dominio asociado al corpus en estudio. El corpus estudiado estuvo compuesto de 1841 documentos científicos de la biblioteca digital de la Comisión Nacional de Energía Nuclear (CNEN) en Brasil, que abordaban varias áreas en el ámbito nuclear. Estos documentos estaban inicialmente en formato PDF y fueron convertidos al formato de texto para su análisis correspondiente. Los resultados confirmaron que la minería de textos es una excelente herramienta para la extracción de conocimientos encapsulados en grandes colecciones de documentos, así como también sirve de apoyo a la gestión de las actividades documentales de investigación en el área nuclear. Pérez; Haro e Saquicela (2017), describen la implementación de técnicas de Bibliomining y Text

Revista ACB: Biblioteconomia em Santa Catarina, Florianópolis (Brasil) - ISSN 1414-0594

Esta obra está licenciada sob uma Licença Creative Commons.

Mining en el Centro de Documentación Regional "Juan Bautista Vázquez" de la Universidad de Cuenca, en el Ecuador. Buscaron analizar las similitudes entre los títulos que buscan los estudiantes y los títulos del material bibliográfico que tomaron en préstamo; para tanto usaron los algoritmos y librerías de la herramienta R, para crear nubes de palabras generadas a partir de términos de búsquedas y los títulos de los libros. Los resultados obtenidos fueron presentados en nubes de palabras, de un total de 100 términos (50 de ellas que conformaron las nubes de palabras de los préstamos y 50 la nube de las búsquedas), 23 términos son comunes tanto en la nube de préstamos como en las búsquedas, obteniendo que existen un comportamiento similar entre los títulos buscados y el material bibliográfico prestado con un porcentaje cercano al 50%.

Nie y Sun (2017) utilizaron las herramientas de extracción de minería de textos para identificar las mayores las actividades académicas y detectar las tendencias de la investigación en diseño en un periodo de 12 años, 2004 a 2015. Describen cómo un abordaje de dos dimensiones de la minería de textos, incluyendo el análisis bibliométrico y el análisis de las redes pueden detectar las tendencias de investigación desde diferentes perspectivas. El enfoque bibliométrico evalúa las tendencias de la producción académica y las tendencias de desarrollo de la investigación en el área de diseño. El análisis de redes intenta encontrar las tendencias de la investigación en cada una de las ramas académicas de investigación y la evolución del núcleo central de los temas de investigación en el área de diseño. Parlina; Kalamullah e Murfi (2020), con el objetivo capturar la estructura científica y la evolución del asunto "big data" utilizando métodos de análisis bibliométricos y basados en minería de textos, recopilaron datos bibliográficos de artículos de revistas publicados entre 2009 y 2018 indexados en la base de datos Scopus. Encontraron un crecimiento significativo de las publicaciones desde 2014. En relación con los principales temas de investigación en las publicaciones sobre big data, agruparon las palabras clave y cada grupo fue etiquetado como un tema. De esa forma los trabajos se dividieron en cuatro subperíodos para observar la evolución temática. El mapeo de temas reveló que la investigación sobre big data está dominada por el análisis que cubren métodos, herramientas, infraestructura de soporte y aplicaciones. Otros aspectos críticos son seguridad y privacidad.

3 MATERIAL Y MÉTODOS

El autor de este trabajo actualiza una base de datos bibliográfica en Endnote X8, desde 1973 hasta 2020. Esta base de datos especializada ha estado en construcción permanente por un periodo de más de diez años. La forma de recolección de los datos bibliográficos ha sido explicada en diversos documentos publicados (URBIZAGASTEGUI & RESTREPO, 2017; URBIZAGASTEGUI & RESTREPO, 2020, URBIZAGASTEGUI & RESTREPO, 2020). Los textos se seleccionaron de esta Base de Datos sobre Bibliometría Brasilera (BB), pero se seleccionaron solo 1 texto de cada año para con este texto hacer la prueba del análisis de textos con el paquete **tm** de R que es simplemente un entorno computacional que puede realizar cálculos numéricos, procesamiento de cadenas con caracteres y mucho más. Para seleccionar cada texto se usó la estrategia de muestra randómica aleatoria sin reemplazos, pero usando los algoritmos pertinentes de R (R Core Team, 2013). Este algoritmo enumera los documentos de 1 a n. Por ejemplo, si en un determinado año se publicaron apenas 100 documentos, entonces el algoritmo enumera los documentos de 1 a 100; luego de estos 100 documentos, se le solicita a R que genere 1 número aleatorio. Digamos que R genera como número aleatorio el 69, entonces el documento 69 es aleatoriamente seleccionado como muestra. De cada uno de estos documentos seleccionados aleatoriamente se copiaron los resúmenes y se llevaron a Notepad. Para guardar cada resumen se usó el

Revista ACB: Biblioteconomia em Santa Catarina, Florianópolis (Brasil) - ISSN 1414-0594

Esta obra está licenciada sob uma Licença Creative Commons.

apellido en ciertos casos abreviando el apellido del autor seguido del año de publicación del documento. Cada documento se guardó en el formato xx.csv, por ejemplo: Azevedo-2020.csv.

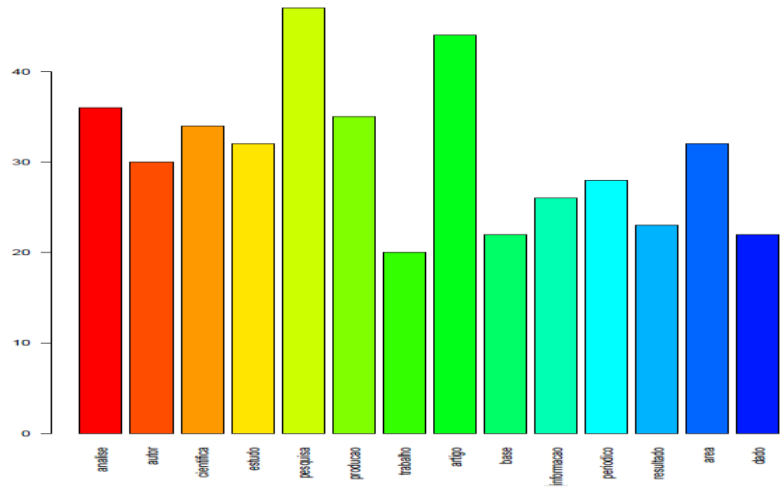
Toda la paquetería de R está pensada y construida para trabajar con documentos en inglés que carecen de acentos. Por lo tanto, para evitar problemas de lectura adecuada, cuando se trabaja con otros idiomas que tienen acentos, se deben remover los acentos, las puntuaciones y convertirlos a minúsculas. La remoción de acentos, puntuaciones y conversión a minúsculas de los textos se hicieron en Notepad.

El análisis de los datos recolectados fue realizado con la ayuda de R (R Core Team, 2013) y los paquetes tm (FEINERER; HORNIK, 2017), NLP (HORNIK, 2021), SnowballC (BOUCHET-VALAT, 2020), cluster (MAECHLER, et al. 2019), topicmodels (HORNIK; GRÜN, 2020), igraph (CSARDI; NEPUSZ, 2006), stats (R Core Team, 2013), gglot2 (WICKHAM et. al., 2020). En estos softwares se realizaron las estadísticas descriptivas e inferenciales pertinentes y se trazaron los gráficos correspondientes. En la investigación se aplicaron algoritmos de agrupación de textos a los artículos académicos escritos en portugués brasileño; esos algoritmos provienen de la tecnología de la información que se implementaron en softwares de libre acceso y se han probado en varios idiomas.

4 RESULTADOS

El paquete tm de R se ejecutó utilizando los resúmenes de los 48 documentos de la muestra que produjo 1424 términos diferentes, con una condición de dispersión del 96% y la palabra de mayor longitud tenía 18 letras. Para identificar las palabras más frecuentes se creó una estructura de datos en la forma de una matriz en la que las filas representan los documentos y las columnas representan los términos más frecuentes. Los valores representan la frecuencia con la que aparece cada palabra en cada documento. En las estructuras semánticas subyacentes de los textos no todos los términos son igualmente informativos, pues algunos términos son inútiles para este propósito. La **Figura 1** representa las palabras utilizadas con un mínimo de 20 veces en los 48 resúmenes de los textos. La figura es auto explicativa.

Figura 1 - representación de las palabras utilizadas con un mínimo de 20 veces



Elaborado por el autor

Revista ACB: Biblioteconomia em Santa Catarina, Florianópolis (Brasil) - ISSN 1414-0594

Esta obra está licenciada sob uma Licença Creative Commons.

Antes de ajustar un modelo, para fines de interpretación y cálculo, se debe eliminar algunas palabras inútiles de la matriz documentos-términos (MDT). Se puede filtrar las palabras basadas en esta información, y seleccionar solo los términos con la puntuación más elevada; luego se aplica la función que remueve los términos muy dispersos para retener la menor cantidad de términos posibles, pero más frecuentes y con mayor valor informativo. Todo este proceso fue realizado con el paquete **tm** (FEINERER; HORNIK, 2017) de R. El valor del argumento de dispersión adoptado fue 0.75 para retener más palabras para la clasificación que con un valor de dispersión menor del elegido que fue 20 palabras. Cuando se aplica la función de eliminación de términos dispersos, el número de términos se reduce. La Tabla 1 muestra las palabras identificas con una frecuencia mayor a 10 ocurrencias.

Tabla 1 - Palabras con frecuencias mayor a 10

Palabra	Frecuencia	Palabra	Frecuencia
pesquisa	47	periodo	14
artigo	44	ciência	14
análise	36	publicado	14
produção	35	maior	13
cientifica	34	ensino	13
estudo	32	relação	13
área	32	indicador	13
autor	30	publicação	12
periódico	28	pais	12
informação	26	documento	12
resultado	23	objeto	12
base	22	obsolescência	12
dado	22	objetivo	11
trabalho	20	brasileiro	11
literatura	18	programa	11
uso	16	citação	11
conteúdo	16	desenvolvimento	11
conhecimento	15	índice	11
gestão	15	grupo	11
pesquisador	15	universidade	11
processo	15	número	11
orçamento	15	modelo	10
ano	14	avaliação	10
lei	14	saúde	10

Elaborado por el autor.

Entre las palabras con frecuencia mayor a 10 no aparece la palabra bibliometría, pero si palabras cercanas a ella o como formas de expresión de la preocupación del campo de la bibliometría: “pesquisa”, “artigo” “análise”, “produção” “científica” y otras. La **Tabla 2** muestra la correlación estadística de Pearson de las tres palabras más frecuentes con las palabras asociadas a cada una de ellas. Estas

Revista ACB: Biblioteconomia em Santa Catarina, Florianópolis (Brasil) - ISSN 1414-0594

Esta obra está licenciada sob uma Licença Creative Commons.

asociaciones se establecieron en un nivel mínimo del 40%. Por ejemplo, la palabra “pesquisa” está asociada a la palabra “área” a un nivel del 56% de correlación; a “instituição” a un nivel del 53% y así sucesivamente a mudança (53%). la palabra “artigo” está más fuertemente asociada a las palabras número, instituição, tendência, país, publicado y autor por encima del 66%. Finalmente, la palabra “Produção” está asociada a organização, autoria, docente, relativa e idioma por encima del 43%; em otras palabras esta asociación es significativa. Esta correlación fue ejecutada con el paquete stats de R (R Core Team, 2013).

Tabla 2 - Correlación de las tres palabras con mayor frecuencia

pesquisa	correlação	artigo	correlação	produção	correlação
área	0.56	número	0.83	organização	0.50
instituição	0.53	instituição	0.82	autoria	0.50
mudança	0.53	tendência	0.72	docente	0.47
longo	0.49	pais	0.71	relativa	0.46
verificou-se	0.49	publicado	0.68	idioma	0.43
necessidade	0.44	autor	0.66		
artigo	0.43	internacional	0.64		
tempo	0.43	longo	0.64		
vinculado	0.43	origem	0.64		
temática	0.42	primeiro	0.62		
concentração	0.41	tipo	0.62		
brasil	0.40	analisada	0.62		

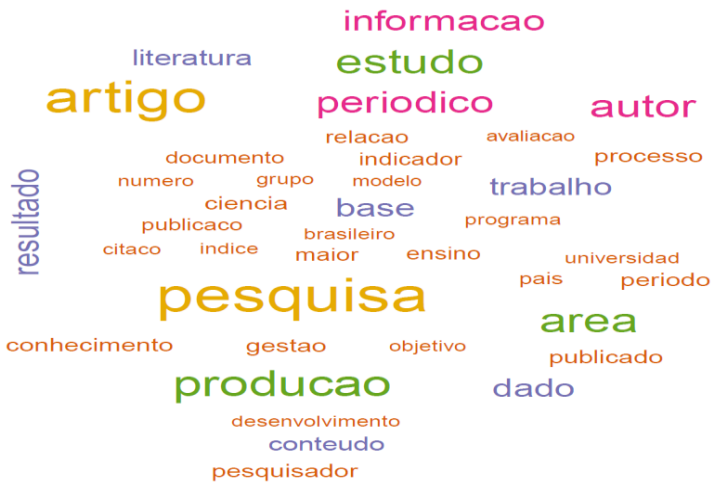
Elaborado por el autor.

La **Figura 2** representa la nube de palabras presentes en el corpus textual con una frecuencia mínima de 20 veces en los 48 resúmenes de los textos analizados. esta nube de palabras fue construida con el paquete wordcloud de R (FELLOWS, 2018).

Figura 2 - Nube de palabras utilizadas con un mínimo de 20 veces

Revista ACB: Biblioteconomia em Santa Catarina, Florianópolis (Brasil) - ISSN 1414-0594

Esta obra está licenciada sob uma Licença Creative Commons.



Elaborado por el autor.

Una situación que se enfrenta comúnmente es el situarse frente a una enorme colección de documentos, pero sin tener ni idea sobre lo que tratan; por lo tanto, uno de los primeros impulsos es organizar y agrupar los documentos según los asuntos abordados. Esto es lo que hacen los sistemas de catalogación y clasificación de documentos en el campo de la bibliotecología y ciencia de la información. El modelado de asuntos se ocupa de clasificar automáticamente conjuntos de documentos en temas similares. El algoritmo Latent Dirichlet Allocation (LDA) que utiliza matemáticas complejas es el que ayuda en estos procesos de clasificación automática de los textos. El supuesto básico del algoritmo LDA es que cada uno de los documentos de una colección consiste en una mezcla de asuntos presentes en toda la colección; es decir, las palabras y no los asuntos son parte de la estructura oculta de los documentos. Asume que existe un número fijo de temas o categorías que se distribuyen sobre los documentos de toda la colección y que cada documento del corpus trata varios temas y a cada término le asigna una probabilidad de pertenecer a un tema en particular. Por lo tanto, dado los documentos y sus estructuras de palabras, lo que se busca es deducir la estructura latente de los asuntos. El algoritmo LDA realiza esta deducción mediante la recreación de los documentos en el corpus a través del ajuste relativo de forma iterativa de la importancia de los temas en los documentos y las palabras en los temas. Latent Dirichlet Allocation (LDA) es un modelo probabilístico generativo que permite que conjuntos de observaciones puedan ser explicados por grupos no observados que identifican algunas partes de los datos que son similares. Por ejemplo, si las observaciones son palabras en documentos, presupone que cada documento es una mezcla de un pequeño número de categorías y la aparición de cada palabra en un documento se debe a una de las categorías a las que el documento pertenece. LDA es un modelo bayesiano jerárquico de tres niveles, en el que cada elemento de una colección se modela como una mezcla finita sobre un conjunto subyacente de temas. Cada tema, a su vez, se modela como una mezcla infinita sobre un conjunto subyacente de probabilidades de temas. En el contexto del modelado de textos, las probabilidades de los temas proporcionan una representación explícita de un documento (BLEI; NG; JORDAN, 2003). En este contexto la **Tabla 3** muestra la agrupación de los 48 documentos estudiados en 5 grupos de tópicos comunes proporcionados por el algoritmo LDA. Los 48 documentos de la muestra fueron agrupados en 5 tópicos diferentes con la identificación de las palabras más recurrentes en cada tópico de agrupación. Por ejemplo, el tópico 1, agrupa a los documentos que tienen las palabras

Revista ACB: Biblioteconomia em Santa Catarina, Florianópolis (Brasil) - ISSN 1414-0594

Esta obra está licenciada sob uma Licença Creative Commons.

informação, conteúdo, literatura, indicador, uso y **análise,** mucho más recurrentemente que en los demás documentos; por lo tanto, estos documentos tienen alta probabilidad de ser similares. Están pues emparentados por su contenido. Estas palabras proceden de los autores de los documentos listados en la **Tabla 4** como: **Faucomp-1995, Fernandes-1999, Figueiredo-1973, Fiuza-1978, Gamboa-2003, Kondo-1998, Oberhofer-1991, Oberhofer-1993** y **Silva-1990**.

Tabla 3 - Tópicos de los documentos de la muestra

	Tópico 1	Tópico 2	Tópico 3	Tópico 4	Tópico 5
1	informação	gestão	cientifica	produção	artigo
2	conteúdo	processo	periódico	trabalho	pesquisa
3	literatura	lei	base	ano	estudo
4	indicador	período	dado	conhecimento	autor
5	uso	tema	pesquisador	orçamento	maior
6	análise	bibliográfica	área	país	

Elaborado por el autor.

La **Tabla 4** muestra la agrupación de los autores de los 48 documentos estudiados que serían más similares según los tópicos estudiados o analizados y mostrados en la Tabla 3. Nuevamente, por ejemplo, el **tópico 4**, agrupa a los documentos que tienen las palabras **produção, trabalho, ano, conhecimento, orçamento** y **país,** con mucha más frecuencia que los demás documentos; por lo tanto, estos documentos también tienen alta probabilidad de ser similares. Están pues emparentados por su contenido. Estas palabras proceden de los autores de los documentos listados en la **Tabla 4** como: **Bienert-2015, Caldas-1980, Caldeira-1975, Caldeira-1979, Chiara-1992, Leite-2008, Mugnai-2009, Oliveira-1984, Pacheco-2001, Robredo-1982, Targino-1987, Witter-2005.**

Tabla 4 - Agrupación de los documentos según los tópicos comunes

Autores de los Documentos	Tópico	Autores de los Documentos	Tópico	Autores de los Documentos	Tópico	Autores de los Documentos	Tópico	Autores de los Documentos	Tópico
Faucomp-1995	1	Azevedo-2020	2	Bufrem-2000	3	Bienert-2015	4	Bernardo-2016	5
Fernandes-1999	1	Guedes-1994	2	Caldeira-1976	3	Caldas-1980	4	Campos-2012	5
Figueiredo-1973	1	Kremer-1989	2	Dias-2019	3	Caldeira-1975	4	Nassif-2018	5
Fiuza-1978	1	Noronha-2002	2	Gomes-1974	3	Caldeira-1979	4	Pellegrini-1977	5
Gamboa-2003	1	Oliveira-1985	2	Lellis-1997	3	Chiara-1992	4	Pinheiro-2011	5
Kondo-1998	1	Peleias-2013	2	Mello-1996	3	Leite-2008	4	Rangel-1981	5
Oberhofer-1991	1	Pinheiro-1983	2	Menegh-2006	3	Mugnai-2009	4	Vincha-2017	5
Oberhofer-1993	1	Ribeiro-2010	2	Rosa-2004	3	Oliveira-1984	4		
Silva-1990	1			Soares-1988	3	Pacheco-2001	4		
				Solano-2014	3	Robredo-1982	4		
				Urbi-2007	3	Targino-1987	4		
				Velho-1986	3	Witter-2005	4		

Elaborado por el autor.

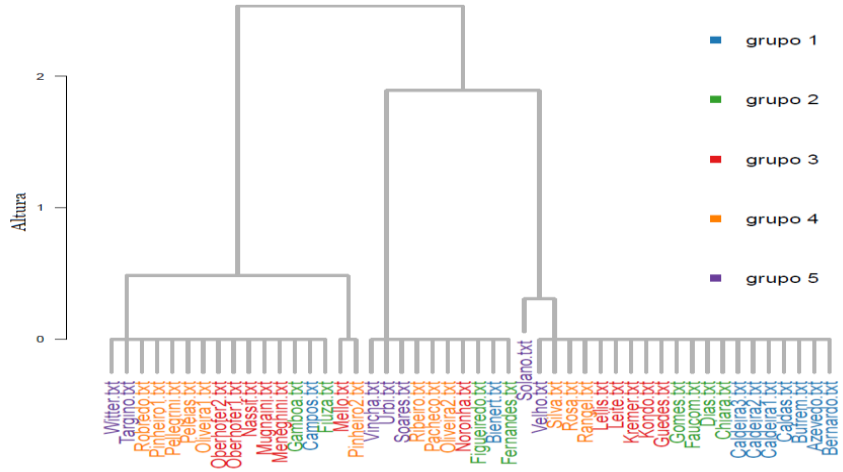
El análisis clúster es un conjunto de técnicas multivariantes utilizadas para clasificar a un conjunto de documentos en grupos homogéneos. La idea es agrupar un conjunto de observaciones en un número dado de clústeres o grupos. Este agrupamiento se basa en la idea de distancia o similitud entre

Revista ACB: Biblioteconomia em Santa Catarina, Florianópolis (Brasil) - ISSN 1414-0594

Esta obra está licenciada sob uma Licença Creative Commons.

las observaciones. La obtención de dichos clústeres depende del criterio o distancia considerados; es decir, depende de lo que consideremos como similar. Para eso se utilizan algoritmos que infieren el número y componentes de los clústeres más aceptables. En la práctica, no se pueden examinar todas las posibilidades de agrupar los documentos y una solución se encuentra en los llamados métodos jerárquicos aglomerativos: se comienza con los objetos o individuos de modo individual; de esa manera se obtienen clústeres iniciales como objetos. Luego se van agrupando de modo que los primeros en hacerlo son los más similares y al final, todos los subgrupos se unen en un único clúster. Una agrupación más detallada es proporcionando por el dendrograma mostrado en la **Figura 3**.

Figura 3 - Dendrograma de agrupación de los documentos más similares entre si



Elaborado por el autor.

Este dendrograma es el resultado de la aplicación de un clúster jerárquico aglomerativo que sirve para encontrar similitudes entre los documentos acomodándolos en grupos, de tal manera que los grupos se encuentren bien separados pero los documentos similares dentro de los grupos están lo más cercano posible. En este caso, el paquete **Cluster** de R (MAECHLER, 2019) realizó varios procedimientos: el primero fue calcular las distancias entre todos los pares de documentos, asumiendo que cada documento constituye un clúster. En el siguiente paso busca dos clústeres más cercanos que los junta y constituye como si fueran uno solo, el proceso se repite hasta que no quedan pares de comparación. El resultado es un dendrograma el cual también puede ser visualizado según las membresías de los documentos. Este dendrograma se creó usando una partición final de 5 conglomerados, lo cual ocurre a un nivel de similitud de aproximadamente 4.0. El primer conglomerado (extremo derecho) se compone de nueve observaciones (las observaciones marcadas de color azul en las filas 42 al 48). El segundo conglomerado, inmediatamente a la izquierda, se compone de 8 observaciones marcadas de verde (las filas 38 al 41). El tercer grupo se compone de 12 observaciones señaladas con el color rojo (las observaciones se dispersan en el dendrograma). El cuarto conglomerado, en el extremo izquierdo, se compone de 12 observaciones (marcadas con el color naranja). El quinto conglomerado que se inicia en el extremo izquierdo, se compone de 7 observaciones (marcadas con el color morado).

Un análisis de clústeres o análisis de grupos o análisis de conglomerados es un método de agrupamiento estadístico utilizado para analizar datos cuantitativos. Normalmente, los datos en observación se dividen en diferentes grupos (clústeres) y se comparan en base a sus características

Esta obra está licenciada sob uma Licença Creative Commons.

específicas. El objetivo de este tipo de análisis es crear grupos homogéneos a partir de objetos individuales heterogéneos. El análisis de clústeres

> es un método estadístico multivariante de clasificación automática de datos. A partir de una tabla de casos-variables, trata de situar los casos (individuos) en grupos homogéneos, conglomerados o clústeres, no conocidos de antemano, pero sugeridos por la propia esencia de los datos, de manera que individuos que puedan ser considerados similares sean asignados a un mismo clúster, mientras que individuos diferentes (disimilares) se localicen en clústeres distintos (FUENTE FERNÁNDEZ, 2001, p. 1).

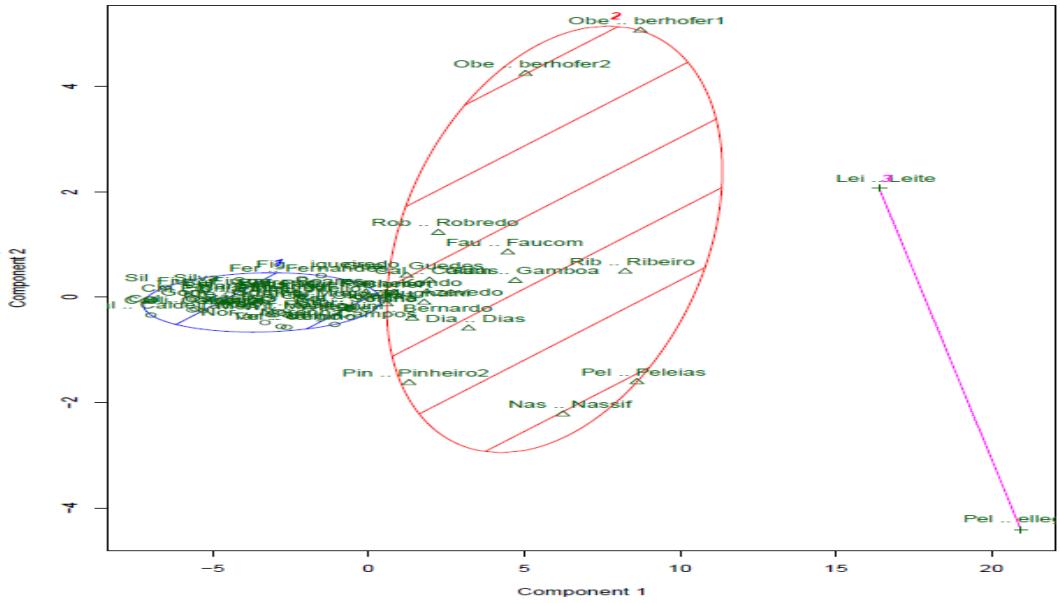
Es decir, es una técnica estadística multivariante cuya finalidad es clasificar un conjunto de datos objetivos en grupos, de forma que los perfiles de los objetivos en un mismo grupo sean muy similares entre sí, por lo tanto, existirá mayor cohesión interna del grupo; por otro lado, si los perfiles de clústeres diferentes son distintos, entonces existirá un aislamiento externo del grupo. Si la clasificación es exitosa, los objetos dentro del clúster estarán muy cercanos unos a los otros en la representación geométrica, y los clústeres diferentes estarán muy apartados. Se trata de una técnica creada para la clasificación de observaciones en grupos. Se persigue reunir las observaciones en grupos lo más homogéneamente posible, de manera que los elementos integrantes de los conglomerados sean muy parecidos. Al mismo tiempo, se busca la máxima heterogeneidad entre clústeres (ALDÁS; URIEL, 2017).

Esta técnica estadística del análisis de conglomerados se aplicó a los asuntos de la muestra de 48 documentos seleccionados de la bibliometría brasileña. Se optó por el método jerárquico pues en este tipo de método los grupos se van formando de forma progresiva, uniendo o separando grupos en función de su similitud. A medida que se avanza en el proceso de agrupamiento se va desarrollando una estructura en forma de árbol, a partir de la cual se decide el número de grupos a obtener, ya que es algo que no se conoce a priori. Para manejar adecuadamente las variables fue necesario homogeneizar sus valores tomando la media valor 0 y la desviación típica 1. Respecto del algoritmo de agrupamiento se decidió emplear el método de Ward para enlaces completos, debido a que proporciona clústeres más homogéneos. Los datos fueron estimados con los paquetes **cluster** del Proyecto R (MAECHLER et. al, 2019). La **Figura 4** muestra la agrupación de los documentos más similares entre sí en tres clústeres más homogéneos.

Figura 4 - Agrupación de los documentos más similares entre sí en tres clústeres

Revista ACB: Biblioteconomia em Santa Catarina, Florianópolis (Brasil) - ISSN 1414-0594

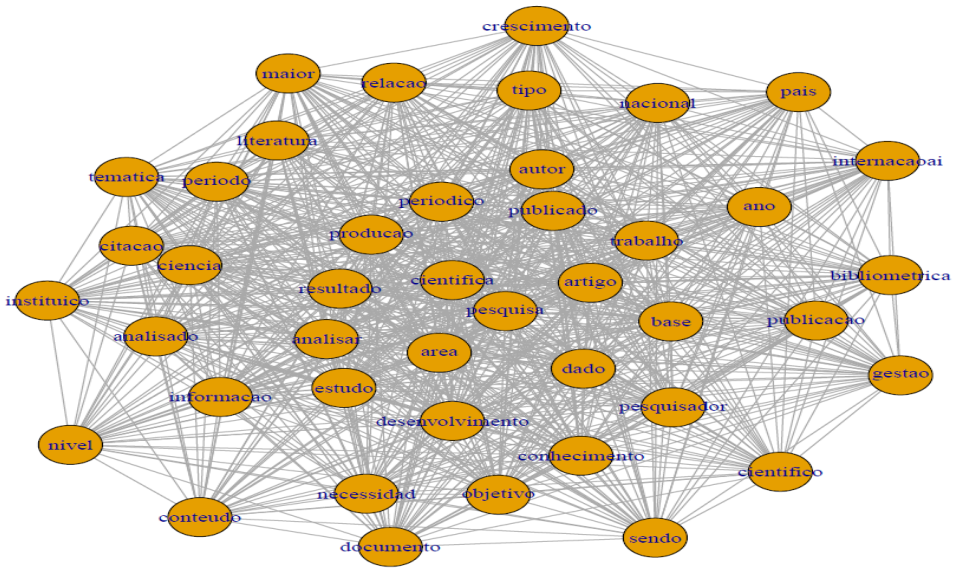
Esta obra está licenciada sob uma Licença Creative Commons.



Elaborado por el autor.

La **Figura 5** muestra la red de relaciones de las palabras o tokenes de los 48 documentos analizados es este trabajo.

Figura 5 - Red de relaciones de las palabras ligadas a la bibliometría brasileña



Elaborado por el autor.

Revista ACB: Biblioteconomia em Santa Catarina, Florianópolis (Brasil) - ISSN 1414-0594

Esta obra está licenciada sob uma Licença Creative Commons.

Para trazar la red se utilizaron solamente las palabras con frecuencias iguales o mayores a 10 usos. Por primera vez aparece la palabra "bibliométrica" pero no bibliometría. Palabras como **"analisar"** y **"analisado"** aparecen por el deficiente trabajo del paquete SnowBallC en el proceso de lematización para el idioma portugués. Sin embargo, la identificación de tokens relativos a las preocupaciones de la investigación en bibliometría brasilera es evidentes. No hay como dudar que en el centro de la red se coloquen palabras como "pesquisa", "científica" y "artigo", que están comprometidas o explicitan las preocupaciones de las investigaciones bibliométricas. Igualmente, las palabras que bordean el centro de la red como "produção", "periodo", "autor", "publicado" y "trabalho", con certeza hacen parte de las mismas exploraciones bibliométricas en el Brasil. Estas mismas observaciones pueden hacerse extensivas a todas las palabras representadas en esta red.

5 DISCUSIÓN Y CONCLUSIÓN

Un problema con los algoritmos de derivación que utilizan la eliminación de afijos y sufijos es que son extremadamente dependientes de los idiomas para los que fueron escritos, ya que se basan directamente en las reglas de formación de palabras de cada idioma para detectarlos y eliminarlos (HONRADO, et. al., 2000). La mayoría de estos algoritmos fueron desarrollados para lidiar con el idioma inglés y no funcionan adecuadamente cuando se aplica a lenguas derivadas del latín como el portugués. El idioma portugués presenta inflexiones que generalmente provocan cambios en la radicalización de las palabras, impidiendo su proceso de conflagración, en el que se basa la derivación. Por este motivo, es necesario utilizar un lematizador especialmente diseñado para el procesamiento de palabras escritas en portugués, pero este lematizador en portugués aún no ha sido desarrollado.

La lematización o stemming es la transformación en lemas de todos los términos que componen el corpus textual, entendiendo por lema la forma de representación de todas las formas flexionadas de una misma palabra. Es una técnica que reduce la complejidad del idioma sin pérdida de información severa. Uno de los algoritmos más conocidos es el propuesto por Porter (1980) que elimina las terminaciones morfológicas e inflexiones comunes de las palabras. Este lematizador está disponible en el paquete SnowballC pero su ejecución es aún deficiente para el idioma portugués. El uso de un lematizador es necesario y urgente, pues si las palabras con el mismo lema se tienen en cuenta como si fuesen una única palabra a la hora de contar la frecuencia de cada término, estas serán contadas como términos diferentes a pesar de compartir una raíz común. Esta representación debe tener en cuenta también la sinonimia y la polisemia, así como las diferencias entre palabras singulares y plurales. Las representaciones basadas en conceptos pueden ser procesadas para soportar jerarquías de conceptos muy sofisticadas pero la sinonimia y la polisemia tienen la gran desventaja de que su implementación es muy compleja y dependen del dominio de múltiples conceptos.

Aunque el trabajo de análisis del contenido del texto todavía requiere un esfuerzo humano, las técnicas de minería de textos son capaces de gestionar de forma cada vez más eficaces grandes volúmenes de información. Con ello, se revela información semántica significativa, que puede ser utilizada como metadatos o para resumir el contenido de los documentos disponibles. Mayor exploración en este campo es necesario y recomendable. La preponderancia de los paquetes de R desarrollados para atender la minería de textos en idioma inglés incide negativamente en la ausencia de paquetes equivalentes para lenguas distintas del inglés, lo que dificulta los procesos del análisis de corpus lingüísticos en portugués e idiomas diferentes del inglés. Por lo tanto, sería recomendable la construcción de paquetes lematizadores en portugués que puedan subsidiar los paquetes de análisis de textos con la paquetería de

Revista ACB: Biblioteconomia em Santa Catarina, Florianópolis (Brasil) - ISSN 1414-0594

Esta obra está licenciada sob uma Licença Creative Commons.

R. A bien decir, existen dos paquetes desarrollados para lidiar con el idioma portugués, pero estos paquetes aún son deficientes.

BIBLIOGRAFÍA

ALDÁS MANZANO, J.; URIEL JIMÉNEZ, E. Análisis multivariante aplicado con R. Madrid: Ediciones Paraninfo, 2017.

ALAZMI, A.R.; ALAZMI, A.R. R Data mining and visualization of large databases. *International Journal of Computer Science and Security*, v.6, n.5, p. 295-314, 2012.

AMADIO; W.J.; PROCACCINO, J.D. Competitive analysis of online reviews using exploratory text mining. *Tourism and hospitality management*, v.22, n.2, p.193-210, 2016.

JUNIOR, A.; GOMES, F.J.F.; RODRIGUES, G.R. Categorização automática de artigos científicos da engenharia de produção utilizando métodos de aprendizagem de máquina. In: ENCONTRO NACIONAL DE ENGENHARIA DE PRODUÇÃO: DESENVOLVIMENTO SUSTENTÁVEL E RESPONSABILIDADE SOCIAL: As Contribuições da Engenharia de Produção, 32, 2012, Anais... Bento Gonçalves, RS, Brasil, 15 a 18 de outubro de 2012.

ARANHA, C.; PASSOS, E. A tecnologia de mineração de textos. *Revista Eletrônica de Sistemas de Informação*, v.5, n.2, p. 1-8, 2006.

ATANASSOVA, I.; BERTIN, M.; MAYR, P. Editorial: Mining Scientific Papers: NLP-enhanced Bibliometrics. Frontiers in Research Metrics and Analytics. *Frontiers in Research Metrics and Analytics*, v.4, art.2, p.1-3, 2019.

BLEI, D.M.; NG, A.Y.; JORDAN, M.I. Latent dirichlet allocation. *Journal of machine Learning research*, *v.*3, n.Jan, p.993-1022, 2003.

BOUCHET-VALAT, M. Package 'SnowballC'. R package version 0.7.0, 2020.

BRAGA, F.DR. Extração semiautomática de taxonomia para domínios especializados usando técnicas de mineração de textos. *Ciência da Informação*, v.45, n.3, p.175-186, set./dez. 2016.

CORRÊA, G.N.; MARCACINI, R.M.; REZENDE, S.O. Uso da mineração de textos na análise exploratória de artigos científicos. Relatório Técnico, Instituto de Ciências Matemáticas e de Computação-USP. Sao Carlos, Sao Paulo, ICMC, p. 1-31, 2012.

CSARDI, G.; NEPUSZ, T. Igraph: Network Analysis and Visualization. R package version 1.2.6, 2020.

DON, A.; ZHELEVA, E.; MACHON, G.; TARKAN, S.; AUVIL, L.; TANYA CLEMENT, T.; SHNEIDERMAN, B.; PLAISANT, C. Discovering interesting usage patterns in text collections: integrating text mining with visualization. PROCEEDINGS OF THE SIXTEENTH ACM

Esta obra está licenciada sob uma Licença Creative Commons.

CONFERENCE ON CONFERENCE ON INFORMATION AND KNOWLEDGE MANAGEMENT, pp. 213-222. ACM, 2007.

FEINERER, I.; HORNIK, K. tm: Text Mining Package. R package version 0.7-3. 2017.

FELLOWS, I. wordcloud: Word Clouds. R package version 2.6, 2018.

FUENTE FERNÁNDEZ, S. de la. Análisis conglomerados. Madrid: Facultad de Ciencias Económicas y Empresariales, Universidad Autónoma de Madrid. Recuperado de: http://tinyurl. com/conglomerados (2020).

GOLDSCHIMDT, R.R.; PASSOS, E. Data Mining: Um Guia Prático. Rio de Janeiro: Campus, 2005.

HORNIK, K. NLP: Natural Language Processing Infrastructure. R package version 0.2.1, 2021.

HORNIK, K.; GRÜN, B. Topicmodels: An R package for fitting topic models. R package version 0.2.12, 2020.

LUIZ, A.R.J.D.C.; MARTINS, D.MS. O uso da descoberta de conhecimento em banco de dados para extração e análise de informação: estudo de caso. Caderno de Estudos em Sistemas de Informação, v.2, no.2, p. 1-16, 2015.

MACHADO, A.P. FERREIRA, R.; BITTENCOURT, I.; ELIAS, E.; BRITO, P. y COSTA, E. Mineração de Texto em Redes Sociais Aplicada à Educação a Distância. Colabor@ - Revista Digital da CVA, v. 6, no. 23, p. 1-21, julho de 2010.

MÄCHLER, M.; ROUSSEEUW, P.; STRUYF, A.; HUBERT, M.; HORNIK, K.; STUDER, M.; ROUDIER, P. y GONZALEZ, J. cluster: Finding Groups in Data: Cluster Analysis Extended Rousseeuw et al. R package version 2.1.0., 2019.

NIE, B.; SUN, S. Using text mining techniques to identify research trends: a case study of design research. *Applied Sciences*, v.7, n.4, p. 401-421, 2017.

HONRADO, A.; LEON, R.; O'DONNEL, R. y SINCLAIR, D. A word stemming algorithm for the Spanish language. Proceedings Seventh International Symposium on String Processing and Information Retrieval. SPIRE 2000, A Curuna, Spain, p. 139-145, 2000.

PARLINA, A.; KALAMULLAH, R.; MURFI, H. Theme Mapping and Bibliometrics Analysis of One Decade of Big Data Research in the Scopus Database. *Information*, v.11, n.2, p.69, 2020.

PÉREZ, W.; HARO, V.; SAQUICELA, V. Bibliomining: Aplicación de text mining para descubrir preferencias de usuarios en el Centro de Documentación Regional “Juan Bautista Vázquez”. Maskana, v.8, p.135-144, 2017.

Esta obra está licenciada sob uma Licença Creative Commons.

PEZZINI, A. Mineração de textos: conceito, processo e aplicações. R. *Eletr. do Alto Vale do Itajaí - REAVI*, v.5, n.8, p.01-13, dez., 2016.

PORTER, M.F. An Algorithm for Suffix Stripping, *Program*, v. 14, n.3, p.130-137, 1980.

R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria., 2013.

URBIZAGASTEGUI, R. y RESTREPO ARANGO, C. Crecimiento de la literatura sobre bibliometria, informetria y cienciometria en el Brasil. Revista ibero-americana de ciência da informação, Brasília, v. 10, n. 1, p. 6-31, jan./jun.2017.

URBIZAGASTEGUI, R. y RESTREPO ARANGO, C. Bibliometría brasileira: la difusión de su literatura. Revista ibero-americana de ciência da informação, Brasília, v. 13, n. 1, p. 200-222, jan./abril 2020.

URBIZAGASTEGUI-ALVARADO, R.; RESTREPO-ARANGO, C. La bibliometría brasilera y el modelo de difusión de innovaciones. Ciência da Informação (Online), Brasilia, Maio 2020.

WICKHAM, H.; CHANG, W.; HENRY, E.; PEDERSEN, T. L.; TAKAHASHI, K.; WILKE, C.; WOO, K. H; YUTANI, H. y DUNNINGTON, D. ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics. R package version 3.3.3, 2020.

Revista ACB: Biblioteconomia em Santa Catarina, Florianópolis (Brasil) - ISSN 1414-0594

Esta obra está licenciada sob uma Licença Creative Commons.