UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Comunicaciones en Estadística
Diciembre 2021, Vol. 14, No. 2, pp. 48–66

Modelamiento de tópicos para identificar patrones en la investigación científica del Covid–19

Topic modeling to identify patterns in Covid-19 scientific research

Carolina Luque[a]
cluque2.d@universidadean.edu.co

Juan Rubriche[b]
juanrubriche@usantotomas.edu.co

Jhon Galvis[c]
jj.galvis916@uniandes.edu.co

Juan Sosa[d]
jcsosam@unal.edu.co

Resumen

Presentamos un modelo de tópicos basado en el método asignación latente de Dirichlet (LDA, por sus siglas en inglés) con el objetivo de examinar patrones en la investigación científica del Covid–19 teniendo en cuenta las publicaciones indexadas en la base datos especializada PubMed. Se toman 4928 resumenes científicos publicados durante el primer semestre de 2020. Se ajusta un modelo LDA utilizando dos tópicos. El primer tópico corresponde a factores de riesgo, severidad y mortalidad por infección viral, mientras que el segundo al impacto de las infecciones respiratorias en la salud pública. La clasificación propuesta brinda una visión global sobre las dos tendencias de investigación presentes a la fecha en la que el análisis tiene lugar. Adicionalmente, los resultados señalan que la aplicación de la metodología propuesta provee un camino para direccionar y hacer más eficiente la revisión bibliográfica en el contexto académico.

***Palabras clave*:** Covid–19, modelos de tópicos, asignación latente de Dirichlet, bases biliográficas, PubMed.

Abstract

We consider a topic modeling approach using latent Dirichlet allocation (LDA) methods aiming to examine patterns in the scientific research of Covid-19 using publications indexed in the PubMed database. A total of 4928 scientific abstracts published during the first semester of 2020 are taken into account. An LDA model is fitted using two topics. The first topic corresponds to risk factors, severity,

[a]Profesor asociado. Facultad de Ingeniería, Escuela de Ciencias Básicas. Universidad Ean
[b]Profesor asociado. Facultad de Estadística. Universidad Santo Tomás
[c]Estudiante. Doctorado en Administración. Universidad de los Andes
[d]Profesor asistente. Departamento de Estadística. Universidad Nacional de Colombia, sede Bogotá

and mortality due to viral infection, whereas the second is the impact of respiratory illnesses on public health. Our classification provides a global overview of these two research trends from the moment the analysis takes place. Additionally, our findings suggest that the systematic application of the proposed methodology provides a way to address and make more efficient the bibliographic review in academic contexts.

***Keywords*:** Covid–19, topic modeling, latent Dirichlet allocation, biliographic bases, PubMed.

1. Introducción

Los avances tecnológicos y la disposición de motores de búsqueda avanzados hacen que sea relativamente facíl obtener en poco tiempo una amplia cantidad de información relacionada con un tema específico. El acceso a contenido científico mediante bases de datos bibliográficas en línea, proporciona documentos variados y de gran magnitud, lo que hace complejo generar una idea macro sobre las tendencias en investigación científica en un campo particular del conocimiento (Gulo and Rúbio, 2015; Trueba-Gómez and Estrada-Lorenzo, 2010). Cada vez hay un mayor número de documentos en línea y la capacidad humana es limitada para consultarlos en su totalidad y leerlos con detalle. Bajo este contexto, sintetizar información disponible en bibliotecas digitales, que crecen de manera vertiginosa y abarcan cientos de años, es útil y relevante para la comunidad académica (Blei, 2012; Blei and Lafferty, 2009).

La posibilidad de identificar de manera estructurada, en una colección de documentos, aquellos artículos que contienen ideas similares a un temática de interés es favorable en términos de tiempo y pertinente para conocer el estado actual de la investigación en un dominio específico (Blei and Lafferty, 2009). Disponer de métodos automatizados para organizar, buscar, comprender y gestionar contenidos, permite caracterizar estudios en un campo particular y reconocer aquellas áreas donde los trabajos son amplios e incluso nichos emergentes de investigación científica.

Enfrentarse al acceso de millones de artículos, requiere de herramientas para explorar y depurar grandes colecciones de literatura académica (Blei and Lafferty, 2009; Valdez et al., 2021). En este trabajo se expone la aplicación del modelo asignación latente de Dirichlet (LDA, por sus siglas en inglés; Blei et al., 2003) con el objetivo de examinar patrones en la investigación científica del Covid–19 durante primer semestre del 2020 a partir de la información disponible en la base datos especializada PubMed. Para la consecución de este propósito, identificamos los términos más frecuentes y relevantes en la investigación científica relacionada con el Covid–19 durante el periodo de interés, y determinamos los temas subyacentes (rasgos latentes) en los estudios científicos asociados con esta enfermedad.

Los resultados de éste trabajo son un complemento metodológico y práctico a la caracterización de la investigación científica en el caso particular del Covid–19.

En términos metodológicos, la aplicación de técnicas de minería de textos exhibe un mecanismo que permite comprimir y distinguir patrones en bases de datos bibliográficas especializadas. En términos prácticos y en el contexto de la investigación científica del Covid–19, los hallazgos revelan tendencias en la literatura que destacan líneas de interés para los estudiosos del área.

El modelado de tópicos ya ha sido empleado para explorar patrones en resumes científicos disponibles en PubMed (ver, Wang et al., 2011). En particular, esta metodológia ya ha sido utilizada para identificar patrones en la literatura académica del Covid–19 (ver, Älgå et al., 2020). A diferencia de estos autores nosotros hacemos un mayor énfasis en la metodología estadística que soporta el modelo LDA.

Este documento se estructura como sigue. En la Sección 2, presentamos los referentes teóricos relacionados con el modelo LDA. En la Sección 3, describimos el proceso de recopilación de datos, el pre–procesamiento, la representación del texto, el ajuste del modelo y la selección del número de tópicos. En la Sección 4, reportamos los resultados de la aplicación del modelo en el contexto de la investigación científica del Covid–19. Finalmente, en la Sección 5, presentamos conclusiones y algunas alternativas de investigación futura.

2. Modelamiento de tópicos

Los modelos de tópicos aluden a un conjunto de metodologías de procesamiento de lenguaje natural que utilizan la automatización informática para consolidar y comprimir el contenido de grandes volúmenes de texto (Blei, 2012; Blei et al., 2003). Son una herramienta que permite descubrir la estructura semántica subyacente (no observada, latente u oculta) en una colección no estructurada de documentos mediante un análisis estadístico de los textos originales (Srivastava and Sahami, 2009). Estos modelos operan bajo el supuesto de que en cualquier colección de texto existe un conjunto de temas (tópicos) ocultos que sirven como pilares para estructurar el conjunto de documentos. Los tópicos se detectan a través del análisis de patrones de lenguaje que se revelan en la colección. En algunos modelos de tópicos (e.g., el LDA), el reconocimiento de los rasgos latentes, se logra luego de eliminar términos no informativos, i.e., que no contribuyen a discriminar entre las temáticas que atraviesan los documentos (e.g., preposiciones o pronombres) (Blei, 2012; Dumais, 2004). Desde un punto de vista técnico, éstos modelos establecen una relación probabilística entre los temas latentes no observados y los rasgos de las variables linguisticas observadas (Kumar and Paul, 2016; Richardson et al., 2014).

Los modelos de tópicos se han utilizado para analizar patrones textuales en documentos de diferente naturaleza. Por ejemplo, se han empleado para estudiar contenido de mesajería móvil y electrónica (Jain, 2021), resúmenes científicos (Griffiths and Steyvers, 2004), informes religiosos (Kim et al., 2020), archivos de periódicos (DiMaggio et al., 2013; Pham et al., 2020), publicaciones en redes sociales (Barry et al., 2018; Chen et al., 2016; Ho and Thanh, 2021; McCallum et al., 2005), do-

cumentos legales (Ashihara et al., 2020; Ovádek et al., 2021), canales comerciales de peticiones, quejas y recursos (Bastani et al., 2019), entre otros. Estos estudios señalan que el modelamiento de tópicos es una herramienta metodológica poderosa y vigente, que facilita el reconocimiento de patrones en el uso de palabras, la identificación de conexiones entre documentos que exhiben características similares, y la inspección de estructuras en colecciones de texto no estructuradas. La configuración temática es un camino para explorar y digerir de manera eficiente grandes cantidades de información (Blei, 2012).

En la literatura hay varios modelos de tópicos (ver, Blei, 2012; Srivastava and Sahami, 2009), dentro de éstos se distingue el modelo LDA. Este es un método que no esta ligado únicamente al análisis de texto, tiene aplicaciones en otros dominios como el filtrado colaborativo, recuperación de imágenes y bioinformática, entre otros (para más detalles ver Blei et al., 2003). En este trabajo nos limitamos a su aplicación en el marco del análisis textual.

2.1. Notación y terminología básica

En al misma línea de Blei et al. (2003), usamos el término "colecciones de texto" para aludir a un todo que contiene "palabras", "documentos" y "corpus". Estos últimos, son entidades que se definen formalmente como sigue:

- Una **palabra**, $\boldsymbol{w}$, es la unidad básica de análisis. Esta se define como un elemento de un vocabulario fijo V y se representa matemáticamente como un vector unitario. La v-ésima palabra en el vocabulario es un V-vector $\boldsymbol{w}$ tal que $w^v = 1$ y $w^u = 0$ para $v \neq u$.
- Un **documento**, d, es una secuencia de N palabras. Así, $\boldsymbol{w}_d$ se utiliza para denotar las palabras observadas en el documento d. En tanto, $\boldsymbol{w}_{d,n}$ es la n-ésima palabra en el documento d.
- Un **corpus**, es una colección de D documentos.

2.2. Modelo de asignación latente de Dirichlet

El LDA es un modelo estadístico que permite identificar patrones en la frecuencia con la que ocurren las palabras y conectar documentos que exhiben tendencias análogas. La idea básica detrás del modelo es que los documentos de la colección comparten un mismo conjunto de temas (tópicos). Así, cada documento se representa como una mezcla aleatoria sobre estos tópicos latentes, los cuales a su vez se caracterizan por una distribución sobre las palabras.

En otros términos, el modelo supone que existen temas latentes presentes en cada documento y que cada palabra contribuye a la constitución de uno o más de ellos. En este sentido, se asume que el número de tópicos, K, es fijo y se distribuye sobre todos los documentos en proporciones diferentes (Blei, 2012; Kumar and

Paul, 2016; Srivastava and Sahami, 2009). La interpretación de la distribución de tópicos es el resultado de detectar la estructura oculta (temas, distribución de temas por documento y asignaciones de palabras por documentos y temas) que probablemente generó la colección de documentos observada.

Como se trata de un algoritmo no supervisado, i.e., que no tiene noción previa de los tópicos que subyacen a la colección, el propósito de su aplicación es descubrir estos temas a través del análisis de los textos originales. En este sentido, el LDA permite organizar el corpus en función de los tópicos detectados. Estos rasgos latentes caracterizan y clasifican los documentos de acuerdo con la similitud temática que hay entre ellos. Bajo esta metodología, no importa el orden de las palabras ni de los documentos (supuestos de intercambiabilidad que ya han sido discutidos por otros autores, e.g., Tian, 2021; Wallach, 2006), dado que, cada texto se considera como una "bolsa de términos" independiente. Conocer las palabras se usaron en el documento y su frecuencia, es suficiente para tomar decisiones sobre los tópicos que subyacen al corpus (Kumar and Paul, 2016).

De acuerdo con Blei (2012), los modelos de tópicos adoptan un proceso generativo (proceso aleatorio mediante el cual se supone surgen las palabras en los documentos) para guiar la complejidad de extraer temas del corpus. En este proceso, se elige aleatoriamente una distribución de temas para cada documento. Seguido, se escoge aleatoriamente un tema de la distribución de tópicos del documento, y finalmente, cada palabra en el documento se extrae del tema seleccionado. De esta manera, el modelo probabilístico asume que los datos son el resultado de un proceso generativo que incluye variables ocultas (i.e., aquellas relacionadas con la estructura de tópicos).

En términos formales, el modelo LDA se distingue por la distribución de tópicos y la distribución sobre el vocabulario. La distribución de tópicos se caracteriza por $\boldsymbol{\theta}_d$ un vector de proporciones para el documento d-ésimo sobre todos los posibles tópicos K. La componente $\theta_{d,k}$ corresponde a la proporción del tópico k en el documento d. De esta manera, el modelo estadístico refleja la intuición de que los documentos presentan múltiples temas y éstos se exhiben en diferentes porcentajes. Luego, para cada texto hay un vector de proporciones de tópicos $\boldsymbol{\theta}_d \mid \boldsymbol{\alpha} \sim \mathbf{Dir}(\boldsymbol{\alpha})$, donde $\mathbf{Dir}(\boldsymbol{\alpha})$ denota la distribución Dirichlet V–dimensional con parámetro vectorial $\boldsymbol{\alpha}$.

Por otro lado, la distribución sobre el vocabulario se caracteriza por $\boldsymbol{\beta}_k$ un vector de probabilidades sobre todas las posibles palabras V. Para cada tópico k hay una distribución sobre las palabras $\boldsymbol{\beta}_k \mid \eta \sim \mathbf{Dir}(\eta)$, donde $\mathbf{Dir}(\eta)$ denota una distribución de Dirichlet simétrica K–dimensional con parámetro escalar η. En tanto, $\boldsymbol{\alpha}$ y η son los hiperparámetros del modelo.

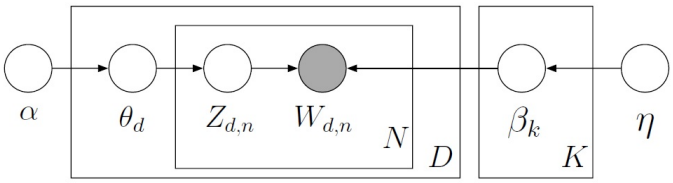


Figura 1: *Modelo gráfico para la asignación latente de Dirichlet (LDA), tomado de Blei (2012). Cada nodo es una variable aleatoria y se etiqueta de acuerdo con su papel en el proceso generativo del corpus. Los nodos ocultos (proporción de tópicos por documento, asignaciones y proporción de tópicos por palabra) no están sombreados. Los nodos observados (palabras en el documento), están sombreados. El rectángulo N denota la colección de palabras en un documento. El rectángulo D la colección de documentos en el corpus. El rectángulo K denota el conjunto de tópicos que atraviesan el corpus.*

Las asignación temática en el documento d-ésimo se denota por $\boldsymbol{z}_d$, donde $z_{d,n}$ es el tópico asignado para la n-ésima palabra en el documento d. Así, a cada término se le asigna un tópico $z_{d,n} \mid \boldsymbol{\theta}_d \sim \mathbf{Mult}(\boldsymbol{\theta}_d)$ donde $z_{d,n} \in \{1, \cdots, K\}$. La n-ésima palabra del documento d, se extrae teniendo en cuenta la distribución de temas por documento y sobre el vocabulario, i.e., $w_{d,n} \mid z_{d,n}, \boldsymbol{\beta}_k \sim \mathbf{Mult}(\boldsymbol{\beta}_{z_{d,n}})$ donde $w_{d,n} \in \{1, \cdots, V\}$. Esto último, muestra que cada palabra en cada documento se extrae de uno de los tópicos, el cual se elige teniendo en cuenta la distribución de tópicos por documento.

En la Figura 1, se ilustra los supuestos detrás del modelo LDA. En esta vía, el proceso generativo bajo el cual se soporta el modelo, conduce a la distribución conjunta de las variables ocultas y observadas (1).

$$p(\boldsymbol{\beta}_{1:K}, \boldsymbol{\theta}_{1:D}, \boldsymbol{z}_{1:D}, \boldsymbol{w}_{1:D}) = \prod_{i=1}^{K} p(\boldsymbol{\beta}_i) \prod_{d=1}^{D} p(\boldsymbol{\theta}_d) \left(\prod_{n=1}^{N} p(z_{d,n} \mid \boldsymbol{\theta}_d) p(w_{d,n} \mid \boldsymbol{\beta}_{1:K}, z_{d,n}) \right) \tag{1}$$

2.3. Inferencia

Indagar sobre la estructura oculta que probablemente generó el corpus observado, conlleva a inferir los tópicos subyacentes a la colección de documentos. En otras palabras, es una cuestión que conduce al problema computacional de utilizar los documentos observados para deducir la configuración latente del corpus. En términos técnicos, esto es capturar la distribución posterior (i.e., la distribución condicional de las variables ocultas dados los documentos), que se expresa como sigue

$$p(\boldsymbol{\beta}_{1:K}, \boldsymbol{\theta}_{1:D}, \boldsymbol{z}_{1:D} \mid \boldsymbol{w}_{1:D}) = \frac{p(\boldsymbol{\beta}_{1:K}, \boldsymbol{\theta}_{1:D}, \boldsymbol{z}_{1:D}, \boldsymbol{w}_{1:D})}{p(\boldsymbol{w}_{1:D})} \tag{2}$$

En la expresión (2), el numerador corresponde a la distribución conjunta especificada en (1). El denominador representa la probabilidad marginal de las observaciones, i.e., la probabilidad de ver el corpus observado bajo cualquier configuración de tópicos. Sin embargo, el número de posibles estructuras temáticas es exponencialmente grande y analíticamente se hace imposible calcular la suma de la distribución conjunta sobre cada posible combinación de estructura de tópicos oculta.

Algunos autores han señalado que la expresión (2) es intratable por métodos analíticos (ver, Blei, 2012; Blei et al., 2003). En tanto, en la literatura se reporta que la distribución posterior debe aproximarse a través de métodos computacionales. Para ello, Blei (2012), propone dos alternativas: algoritmos basados en muestreo y métodos variacionales.

Los algoritmos basados en muestreo proporcionan un mecanismo para generar muestras de la posterior con el fin de aproximarla a través de una distribución empírica. El historial iterativo que se genera mediante estos métodos se utiliza para aprender sobre los parámetros del modelo a posteriori. Blei (2012) afirma que de éstos algoritmos el más común en el modelado de tópicos es el muestreador de Gibbs, una herramienta computacional que hace parte de los métodos de Cadenas de Markov Monte Carlo (MCMC, por sus siglas en inglés). Los detalles teóricos y prácticos para la implementación del muestreador de Gibbs en el contexto del modelado de tópicos se encuentran en Darling (2011).

Por otro lado, Blei (2012) señala que los métodos variacionales son una alternativa a los algoritmos basados en muestreo. Estos no aproximan la distribución posterior a través de muestras sino que convierten el proceso de inferencia en un problema de optimización, postulando una familia parametrizada de distribuciones sobre la estructura oculta, para luego identificar el miembro de dicha familia que tiene una mayor proximidad a la distribución posterior de interés. Para más detalles sobre estos métodos consultar Blei et al. (2003) y Wainwright and Jordan (2008).

3. Metodología

3.1. Recopilación de los datos

PubMed (Public MEDLINE, `https://pubmed.ncbi.nlm.nih.gov/`) es una base de datos especializada de acceso libre que comprede más de 30 millones de citas de literatura biomédica de MEDLINE, revistas de ciencias de la vida y libros en línea. Las citas incluyen enlaces a contenido de texto completo de PubMed Central y sitios web de editor. Es actualizada constantemente por el Centro Nacional de Información Biotecnológica (NCBI, por sus siglas en inglés) de la Biblioteca Nacional de Medicina de EE.UU (NLM, por sus siglas en inglés). Es una de las

bases bibliográficas más conocidas y manejadas por los profesionales en el área de la medicina y ciencias de la vida, dado que recoge bibliografía desde 1950 y tiene un incremento promedio de unas 800 mil referencias al año (Trueba-Gómez and Estrada-Lorenzo, 2010).

Para el desarrollo de este trabajo se toma un conjunto de artículos de PubMed publicados durante el primer semestre de 2020 y relacionados con Covid–19. Para recuperar los registros se utiliza la librería `easyPubMed` (Fantini, 2017) disponible en R, que proporciona una interfaz para acceder y descargar el PMID (número único de registro en PubMed), DOI, título, resumen, fecha de publicación (año, mes, día), nombre de la revista (título, abreviatura), palabras clave y autor (nombres, afiliación, dirección de correo electrónico)[1]. Se extrae información de 4928 artículos [2], de los cuales sólo 3024 hacen parte del análisis dado que cumplen dos condiciones: disponibilidad del resumen y texto en inglés. La información de los artículos se consolida en un archivo *csv* y se selecciona como variable objeto de análisis el resumen de los artículos.

3.2. Pre–procesamiento del texto

Se conforma un corpus con los 3024 resúmenes de los artículos descargados de PubMed. Posteriormente, se realiza el pre–procesamiento mediante la interfaz que proporciona la librería `tm` (Feinerer, 2013).

El pre–procesamiento permite limpiar y estructurar el texto de entrada para posteriormente representarlo sólo con aquellos términos que lo caracterizan (palabras que reflejan el significado central de la escritura, Valdez et al., 2021). El pre–procesamiento incluye: eliminar carácteres especiales, suprimir números y puntuación, transformar a minúsculas, remover texto unicode, reducir espacios en blanco, excluir lenguaje sin importancia semántica (palabras que gramaticalmente son necesarias para la lectura pero que no reflejan el significado central del texto, tales como: preposiciones, pronombres, adjetivos comunes, entre otros), eliminar palabras que corresponden a características generales del texto (e.g., *doi*) o que hacen parte del tema general de los documentos (e.g.,*covid*) y en tanto, se sabe de antemano son un términos recurrentes en todos los documentos.

También se realiza lematización del texto con el objetivo de reducir dimensionalidad. Esto último, es una transformación predeterminada de la librería `tm` y se fundamenta en el algoritmo propuesto por Porter (2006), el cual elimina terminaciones morfológicas e inflexiones comúnes de las palabras permitiendo agrupar bajo una misma raíz aquellas que tienen el mismo significado. De esta manera, palabras como *value* y *valuing* serán parte de la frecuencia del lema *valu*.

[1] El código para la descarga y modelado de los datos esta disponible en `https://github.com/Luque-ZabalaC`

[2] Los datos se descargan el 12 de junio de 2020

3.3. Representación de los datos

Los datos del modelo LDA corresponden a frecuencias de términos, por ello, es necesario hacer la representación del texto a través de la matriz de documentos–términos (DTM, por sus siglas en inglés) constituida por D filas y V columnas. Cada fila de ésta matriz representa un vector de frecuencias de los términos en cada documento y cada columna representa un vector de frecuencias de cada término en los documentos. Se utilizan dos métodos de ponderación en la construcción de la DTM: (i) la frecuencia de los términos y (ii) el esquema "frecuencia de términos–frecuencia de documentos inversa"(TF–IDF, por sus siglas en inglés, para más detalles ver Aizawa, 2003; Qaiser and Ali, 2018). El esquema TF–IDF se utiliza para reducir la dimensionalidad de la DTM, delimitando el vocabulario V a los términos más importantes en el corpus (e.i., permite identificar el conjunto de palabras que discriminan para los documentos de la colección), de esta manera reduce documentos de longitud arbitraria a listas de números de longitud fija (Blei et al., 2003).

La estadística TF–IDF se define para cada término en cada documento como $1 + \log\left(\frac{D}{n_w}\right)$ cuando el término w esta en el documento d y cero en caso contrario (Richardson et al., 2014). Aquí D es el total de documentos en el corpus y n_w es el número de documentos en los cuales aparece la palabra w. La ponderación de TF–IDF aumenta proporcionalmente al número de veces que una palabra aparece en el documento, pero es compensada por la frecuencia de la palabra en la colección de documentos, lo cual ayuda a controlar el hecho de que algunas palabras resulten más comunes que otras y permite filtrarlas.

De esta manera, luego del pre–procesamiento del corpus, quedan sólo aquellos términos que aportan información en relación a las estructuras semánticas subyacentes en los documentos. Además de hacer la selección de los términos más relevantes a través de la estadística TF–IDF, es usual reducir la esparcidad de la matriz DTM (Gulo and Rúbio, 2015), lo cual conduce a tener menos términos pero más comunes. Esto último, favorece la interpretación de los rasgos latentes y el rendimiento computacional.

3.4. Ajuste del modelo

La implementación del modelo se lleva a cabo mediante el paquete `topicmodels` (Grün and Hornik, 2011) disponible en R. Este paquete dispone de dos algoritmos para ajustar un modelo de tópicos: (i) el algoritmo de maximización de expectativa variacional (VEM, por sus siglas en inglés) y (ii) el algoritmo basado en el muestreador de Gibbs. Nosotros optamos por la primera alternativa para inferir la estructura oculta en el corpus.

3.5. Selección del número de tópicos

Un campo de investigación activo en el modelado de tópicos es la evaluación y contraste de modelos (ver, Blei, 2012). En la literatura se presentan diversas alternativas para abordar ésta problemática. Por ejemplo, se hace el contraste usando un número diferente de tópicos o distintos valores de los hiperparámetros del modelo.

Realizar el proceso de estimación teniendo en cuenta un número diferente de tópicos no sólo permite comparar modelos, también contribuye a identificar la cantidad de temas latentes en el corpus. Algunos autores manifiestan que esta vía junto con el criterio de máxima verosimilitud posibilita determinar el número óptimo de tópicos del modelo (Buntine, 2009). En este sentido, hacemos la selección de ésta cantidad ajustando modelos LDA con un número distinto de tópicos. Siguiendo a Buntine (2009), la cantidad óptima se determina cuando la log-verosimilitud se maximiza con respecto a los hiperparámetros $\boldsymbol{\alpha}$ y η. A estos últimos, les asignamos valores iniciales aleatorios en el algoritmo VEM (Grün and Hornik, 2011; Grün et al., 2019).

4. Resultados

A la fecha de recopilación de datos, PubMed cuenta con 17187 publicaciones en diferentes idiomas sobre Covid–19. Por limitaciones en la cantidad de registros que se pueden descargar a través de la interfaz **fetch_pubmed_data** (Fantini, 2017), solamente se obtiene información de 4928 documentos científicos, de los cuales 1904 no disponen de resumen. En tanto, el análisis se realiza con un corpus de 3024 documentos en inglés, donde cada resumen es un documento.

El prepocesamiento del texto sigue las transformaciones usuales (Gulo and Rúbio, 2015; Richardson et al., 2014). Adicionalmente, se hace necesario definir y remover un listado de palabras personalizadas por tratarse de términos generales o recurrentes en el contexto. Antes del pre–procesamiento, el corpus tiene un total de 43795 términos, esta cantidad de palabras se reduce en un 65 % aproximadamente, luego de las transformaciones realizadas sobre el texto. La depuración, permite consolidar una DTM con un número de filas igual al número de documentos seãlado previamente y 15380 columnas (términos). Esta matriz arroja una esparcidad de 100 %, lo cual evidencia que la mayoría de los valores de entrada de la matriz son cero. También se identifica la existencia de términos hasta con 60 caracteres, los cuales por su longitud son palabras sin sentido que aparecen producto de agrupaciones en el proceso de limpieza.

Con el propósito de reducir la dimensionalidad de la DTM, se toman decisiones fundamentadas en el análisis descriptivo de la misma y en pro de descartar aquellos términos que no representan información relevante para el modelo de tópicos. Al examinar el número de caracteres que conforman los términos incluidos en la DTM, se identifica que no hay palabras compuestas por uno o dos caracteres. Aquellas

compuestas por tres caracteres son en total 974 y las compuestas por más de veinte caracteres son 159. En éstos dos últimos casos, se evidencia que en su mayoría corresponden a agrupaciones (de caracteres o palabras) sin sentido y que aparecen en muy pocos documentos. Al revisar la frecuencia con la que ocurren los términos en los documentos, se observa que 7089 de los 15380 aparecen sólo una vez en el corpus. Estos primeros hallazgos, conducen a hacer un primer filtro de la DTM, considerando solamente aquellos términos con frecuencia mayor o igual a 2 y que están compuestos por una cantidad de caracteres entre 4 y 20. Este primer filtro reduce el número de términos a 6517.

Como que hay términos que pueden figurar en muy pocos documentos, se toma la decisión de reducir la esparcidad de la matriz eliminando aquellos que se presentan en 10 o menos documentos. Lo anterior, reduce tanto la cantidad de términos (2426) como la espacidad (98 %) de la DTM. Por último, con el fin de consolidar un vocabulario con los términos más relevantes en la colección de documentos, se utiliza la estadística TF–IDF. El cálculo de ésta medida evidencia que 8 de los 3024 documentos, luego de la depuración de la matriz quedan sin términos (ésto posiblemente ocurre porque algunos resumenes son generales y no muy extensos).

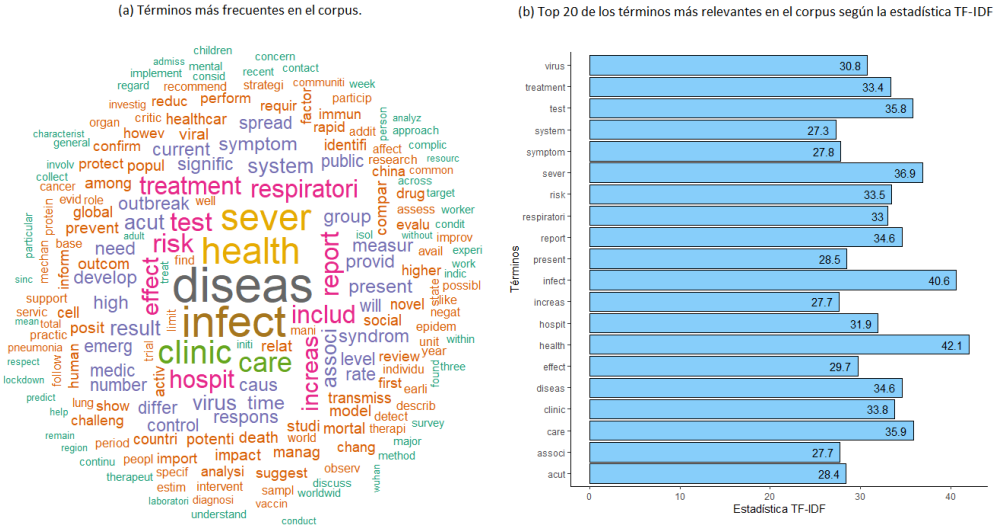
El valor minímo de la estadística TF–IDF es de 0.560 y el valor máximo es 40.615; el valor promedio es de 5.253 con una desviación estándar de 5.295 aproximadamente. En la literatura no se reporta un valor a partir del cual ésta estadística sea óptima, sólo se menciona que lo ideal es recuperar aquellos términos con la puntuación más alta porque ello implica la selección de las palabras más relevantes en la colección de documentos (Gulo and Rúbio, 2015). En este sentido, se decide retener aquellos términos cuyo valor de TF–IDF es superior al primer cuartil (1.880). Dado lo anterior, se reduce la dimensionalidad de la DTM a 3024×1819, con una esparcidad del 98 % y con términos de longitud máxima 18.

En la Figura 2a, se resaltan los términos más frecuentes en el corpus, mediante una nube de palabras. Aquellas palabras que más se destacan en la nube, tienen una frecuencia superior a 900 en la colección de documentos. Por ejemplo, el término *diseas* tiene una frecuencia de 2830, seguido por *infect* con una frecuencia de 2475. En la Figura 2b, se presenta el top 20 de las palabras con mayor importancia dentro del corpus, según la estadística TF–IDF. Al comparar los dos gráficos, se evidencia que términos como *diseas, infect, sever*, entre otros, son frecuentes y relevantes en la colección de documentos.

El ajuste del modelo LDA para diferentes valores de K (Figura 3), señala un valor máximo para la log–verosimilitud cuando se utiliza un número de tópicos igual a 2. Este hallazgo, conduce a elegir dicha cantidad como la opción óptima para modelar los datos de estudio.

En la Figura 4, se muestra el top 20 de los términos que tienen mayor probabilidad de ser generados por cada tópico. Los términos con mayor probabilidad de pertenecer al tópico 1 son: *sever* (1.32 %), *care* (0.93 %), *test* (0.91 %), increas (0.88 %) y *hospit* (0.80 %); los de mayor probabilidad de formar parte del tópico 2 son: *health* (1.61 %), *diseas* (1.46 %), infect (1.32 %) y *clinic* (1.30 %). Al comparar los términos que componen cada tópico (Figura 4) se evidencia que la investigación

Figura 2: Representación gráfica de las palabras más frecuentes y relevantes en el corpus luego de la reducción de dimensionalidad de la DTM.



científica del Covid–19 durante el primer semestre de 2020 se direcciona hacia factores de riesgo, severidad y mortalidad por infección viral (tópico 1) e impacto de las infecciones respiratorias en la salud pública (tópico 2).

En la Figura 4 también se puede evidenciar palabras como *diseas* e *infect* que son comunes a ambos tópicos pero tienen una probabilidad de pertenencia diferente. Esto evidencia que bajo la metodología propuesta, las palabras presentan cierta superposición, revelando el aporte de cada término a la constitución de los tópicos latentes.

En la Figura 5, se muestran las palabras con mayor diferencia en sus probabilidades β de pertenencia al tópico 1 y al tópico 2. Dicha diferencia se calcula usando el logarítmo base 2 de la razón entre β_2 y β_1. El gráfico nos permite visualizar los términos que tienen un mayor chance de pertenecer a un tópico frente al otro. Los términos cuyo logaritmo es negativo tienen una mayor probabilidad de capturar la temática del tópico 1. Mientras que las palabras que tienen asociado un logaritmo positivo tienen mayor chance de capturar la temática del tópico 2. Las palabras como *asymptomat, adult* y *safeti* corresponden a términología relacionada con los factores de riesgo, severidad y mortalidad por infección viral (tópico 1) y las palabras como *sensit, educ* y *therapeut* se asocian con el léxico propio del impacto de las infecciones respiratorias en la salud pública (tópico 2).

Por otra parte, el LDA también permite modelar cada documento como una mezcla de tópicos. Es posible examinar las proporciones de cada tópico en cada documen-

Figura 3: Distribución de la log–verosimilitud según el número de tópicos incluidos en el modelo.

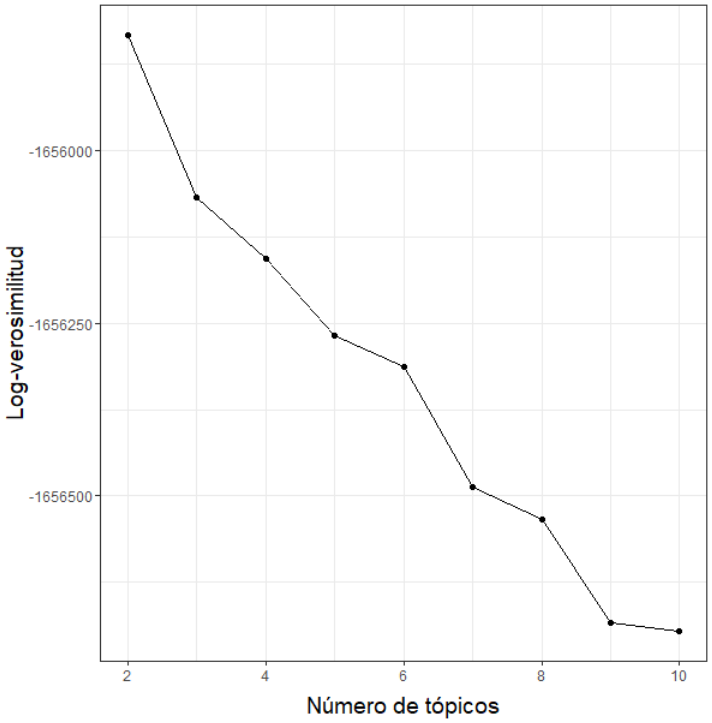


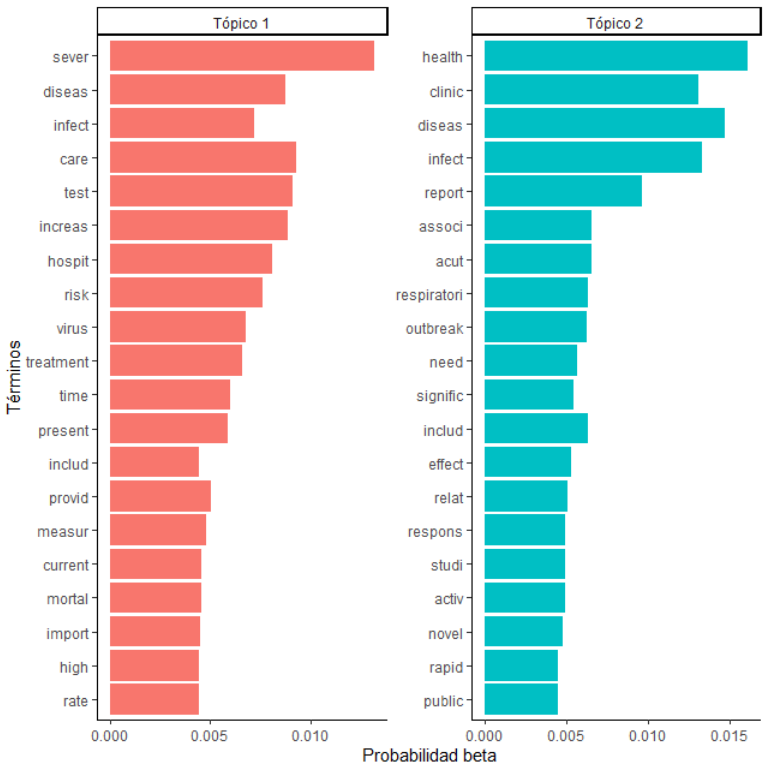
Tabla 1: Distribución θ por tópico para 5 documentos del corpus.

Documento	Tópico 1	Tópico 2
1	0.5216585	0.4783415
2	0.4718221	0.5281779
3	0.5573852	0.4426148
4	0.4826220	0.5173780
5	0.5033839	0.4966161

to. Estas probabilidades se denominan θ y las interpretamos como la proporción estimada de palabras en ese documento que se generan a partir de determinado tópico Silge and Robinson (2017). Por ejemplo, el modelo propuesto estima que cerca del 52.16 % de los términos en el documento 1 son generados en el tópico 1 y 47.83 % de los términos en este mismo documento son generados por el tópico 2 (Tabla 1).

Finalmente, el modelo LDA propuesto clasifica 1469 de los 3024 documentos en el tópico 1 y 1547 en el tópico 2. Los 8 restantes no fueron clasificados por el modelo

Figura 4: Top 20 de las palabras que caracterizan cada tópico.



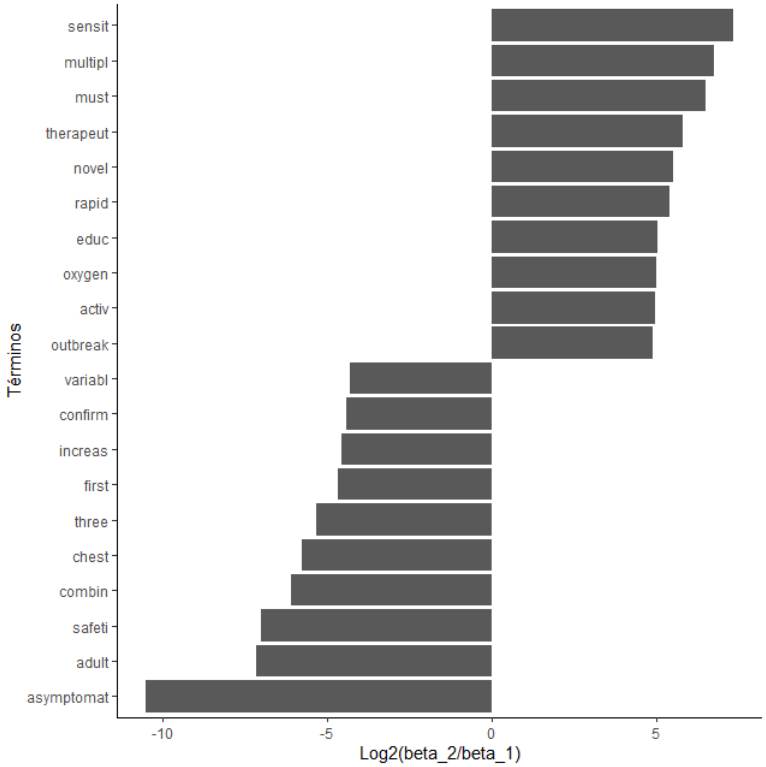
debido a que no se contaba con términos para su clasificación.

5. Conclusiones

A partir de los resultados se llega a las siguientes conclusiones.

- La aplicación del modelo LDA sobre los resúmenes del repositorio PubMed, permite extraer patrones de investigación científica del Covid–19 para el primer semestre del año 2020. Se logra distinguir tópicos latentes que enmarcan dos tendencias en la investigación en ésta área del conocimiento: (i) factores de riesgo, severidad y mortalidad por infección viral e (ii) impacto de las infecciones respiratorias en la salud pública.
- La clasificación propuesta brinda una visión a los investigadores en relación a las tendencias que subyacen en los estudios iniciales sobre el Covid–19. Además facilita la revisión de la literatura siguiendo alguna de las dos co-

Figura 5: Términos con mayor diferencia entre los logaritmos de sus probabilidades β asociadas a cada tópico.



rrientes inferidas. Este estudio pone de manifiesto que la implementación de técnicas de modelamiento de tópicos para revisión de literatura, debe ser considerada para potenciar la búsqueda sistemática de tópicos latentes en los recursos bibliográficos dispuestos en la red.

- La aplicación de minería de textos posibilita caracterizar la colección de documentos a través de las extracción de las palabras más frecuentes y relevantes en el corpus. Palabras como *sever*, *care*, *test*, *increas*, *hospit*, *health*, *diseas*, infect y *clinic* son algunos de los términos que permitien sintetizar el contenido de los resúmenes objeto de análisis y develar los tópicos latentes manifiestos en los documentos.

- El pre–procesamiento del texto es un paso fundamental, que debe estar sujeto a la naturaleza de los documentos que conforman el corpus. Un pre–procesamiento adecuado debe conducir a la conformación de un vocabulario que contenga palabras informativas para el modelo, dado que ésto condiciona la clasificación de los documentos y por ende la conformación y posterior

interpretación de los tópicos.

Son varias las cuestiones alrededor de la metodología propuesta, que pueden orientar el diseño y desarrollo de trabajos futuros. A continuación citamos algunas, relacionadas con el modelamiento de tópicos en general y de manera particular de documentos científicos.

- El análisis que presentamos en este documento se realiza sólo con el resumen de algunos artículos de PubMed. Es de trabajos futuros contrastar los resultados expuestos tomando el texto completo de los artículos, lo cual permitiría examinar sí el resumen es lo suficientemente informativo para hacer una buena clasificación de los documentos. También valdría la pena examinar la aplicación del modelo LDA para todos documentos del repositorio PubMed, i.e., sin restricciones de búsqueda.
- En este trabajo consideramos sólo publicaciones iniciales, teniendo en cuenta el primer semestre del año 2020. Es de investigaciones posteriores considerar un análisis de tópicos dinámico (ver, Blei and Lafferty, 2007) que permita evaluar la evolución de las tendencias en investigación en ésta y otras áreas a lo largo del tiempo.
- En este documento se ajusta el modelo a través del algoritmo de maximización de expectativa variacional (VEM). Es de trabajos futuros contrastar los resultados que arroja la aplicación de este algoritmo con los que se obtienen al implementar el muestreador de Gibbs. También, es de interés posterior examinar las implicaciones que tiene en los resultados, el hecho de postular diferentes valores para los hiperparámetros del modelo LDA.

Recibido:
Aceptado:

Referencias

Aizawa, A. (2003). An information-theoretic perspective of tf–idf measures. *Information Processing & Management*, 39(1):45–65.

Älgå, A., Eriksson, O., and Nordberg, M. (2020). Analysis of scientific publications during the early phase of the covid-19 pandemic: topic modeling study. *Journal of medical Internet research*, 22(11):e21559.

Ashihara, K., El Vaigh, C. B., Chu, C., Renoust, B., Okubo, N., Takemura, N., Nakashima, Y., and Nagahara, H. (2020). Improving topic modeling through homophily for legal documents. *Applied Network Science*, 5(1):1–20.

Barry, A. E., Valdez, D., Padon, A. A., and Russell, A. M. (2018). Alcohol advertising on twitterâa topic model. *American Journal of Health Education*, 49(4):256–263.

Bastani, K., Namavari, H., and Shaffer, J. (2019). Latent dirichlet allocation (lda) for topic modeling of the cfpb consumer complaints. *Expert Systems with Applications*, 127:256–271.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4):77–84.

Blei, D. M. and Lafferty, J. D. (2007). A correlated topic model of science. *The annals of applied statistics*, 1(1):17–35.

Blei, D. M. and Lafferty, J. D. (2009). Topic models. In *Text mining*, pages 101–124. Chapman and Hall/CRC.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Buntine, W. (2009). Estimating likelihoods for topic models. In *Asian Conference on Machine Learning*, pages 51–64. Springer.

Chen, L., Hossain, K. T., Butler, P., Ramakrishnan, N., and Prakash, B. A. (2016). Syndromic surveillance of flu on twitter using weakly supervised temporal topic models. *Data mining and knowledge discovery*, 30(3):681–710.

Darling, W. M. (2011). A theoretical and practical implementation tutorial on topic modeling and gibbs sampling. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 642–647.

DiMaggio, P., Nag, M., and Blei, D. (2013). Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of us government arts funding. *Poetics*, 41(6):570–606.

Dumais, S. T. (2004). Latent semantic analysis. *Annual review of information science and technology*, 38(1):188–230.

Fantini, D. (2017). easypubmed: An r package for search and retrieve scientific publication records from pubmed. Technical report.

Feinerer, I. (2013). Introduction to the tm package text mining in r. Technical report.

Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National academy of Sciences*, 101(suppl 1):5228–5235.

Grün, B. and Hornik, K. (2011). topicmodels: An r package for fitting topic models. *Journal of statistical software*, 40(1):1–30.

Grün, B., Hornik, K., and Grün, M. B. (2019). Package âtopicmodelsâ.

Gulo, C. A. and Rúbio, T. R. (2015). Text mining scientific articles using the r. In *Doctoral Symposium in Informatics Engineering*.

Ho, T. and Thanh, T. D. (2021). Discovering community interests approach to topic model with time factor and clustering methods. *Journal of Information Processing Systems*, 17(1):163–177.

Jain, E. G. (2021). A comparative analyzing of sms spam using topic models. In *Innovations in Information and Communication Technologies (IICT-2020)*, pages 91–99. Springer.

Kim, S.-H., Lee, N., and King, P. E. (2020). Dimensions of religion and spirituality: A longitudinal topic modeling approach. *Journal for the scientific study of religion*, 59(1):62–83.

Kumar, A. and Paul, A. (2016). *Mastering text mining with R*. Packt Publishing Ltd.

McCallum, A., Corrada-Emmanuel, A., and Wang, X. (2005). Topic and role discovery in social networks.

Ovádek, M., Dyevre, A., and Wigard, K. (2021). Analysing eu treaty-making and litigation with network analysis and natural language processing. *Frontiers in Physics*, 9:202.

Pham, Q., Stanojevic, M., and Obradovic, Z. (2020). Extracting entities and topics from news and connecting criminal records. *arXiv preprint arXiv:2005.00950*.

Porter, M. F. (2006). An algorithm for suffix stripping. *Program*.

Qaiser, S. and Ali, R. (2018). Text mining: use of tf-idf to examine the relevance of words to documents. *International Journal of Computer Applications*, 181(1):25–29.

Richardson, G. M., Bowers, J., Woodill, A. J., Barr, J. R., Gawron, J. M., and Levine, R. A. (2014). Topic models: A tutorial with r. *International Journal of Semantic Computing*, 8(01):85–98.

Silge, J. and Robinson, D. (2017). *Text mining with R: A tidy approach*. .ºReilly Media, Inc.".

Srivastava, A. N. and Sahami, M. (2009). *Text mining: Classification, clustering, and applications*. CRC press.

Tian, Y. (2021). A multilayer correlated topic model. *arXiv preprint arXiv:2101.02028*.

Trueba-Gómez, R. and Estrada-Lorenzo, J.-M. (2010). La base de datos pubmed y la búsqueda de información científica. *Seminarios de la Fundación Española de Reumatología*, 11(2):49–63.

Valdez, D., Picket, A. C., Young, B.-R., and Golden, S. (2021). On mining words: The utility of topic models in health education research and practice. *Health Promotion Practice*, 22(3):309–312.

Wainwright, M. J. and Jordan, M. I. (2008). Introduction to variational methods for graphical models. *Foundations and Trends in Machine Learning*, 1:1–103.

Wallach, H. M. (2006). Topic modeling: beyond bag-of-words. In *Proceedings of the 23rd international conference on Machine learning*, pages 977–984.

Wang, H., Ding, Y., Tang, J., Dong, X., He, B., Qiu, J., and Wild, D. J. (2011). Finding complex biological relationships in recent pubmed articles using bio-lda. *PloS one*, 6(3):e17243.