Modelo para la determinación de temas de investigación a partir de la aplicación de técnicas de aprendizaje no supervisado

PROYECTO DE GRADO

Andrés Hernández

Asesor

Javier Díaz Cely

PhD, Computer Science

FACULTAD DE INGENIERÍA

MAESTRÍA EN CIECIA DE DATOS

SANTIAGO DE CALI

2020

Modelo para la determinación de temas de investigación a partir de la aplicación de técnicas de aprendizaje no supervisado

Andrés Hernández

Trabajo de grado para optar al título de
Máster en Ciencia de Datos

Asesor
Javier Díaz Cely
PhD, Computer Science

FACULTAD DE INGENIERÍA
MAESTRÍA EN CIENCIA DE DATOS
SANTIAGO DE CALI
2020

CONTENIDO

LISTA DE TABLAS

LISTA DE FIGURAS

LISTA DE ANEXOS

RESUMEN

Durante el desarrollo de una propuesta de investigación, una de las actividades más importantes, consiste en la recopilación de referencias sobre sobre las cuales se van a generar las bases de la propuesta a desarrollar. Identificar el estado actual del tema de investigación de interés, las principales revistas científicas que lo abordan, y las colaboraciones que actualmente se llevan a cabo, se convierte en una capacidad fundamental para el proceso investigativo y el desarrollo científico.

En el presente trabajo, se formuló una propuesta para abordar el problema que representa, realizar un proceso de recopilación de información (referencias bibliográficas), teniendo en cuenta el crecimiento exponencial actual de la cantidad de información y el método de revisión, que consiste en una depuración manual e individual de las referencias, lo cual impacta directamente el tiempo de desarrollo de la investigación.

La propuesta de este trabajo consiste, en utilizar el algoritmo LDA (*Latent Dirichlet Allocation*) como método para acercarse a la solución del problema de identificación de tópicos, teniendo en cuenta que es el modelo tópico más utilizado, permite una buena interpretación de los datos y existe amplia disponibilidad de literatura referente al tema.

Los elementos usados en el abordaje del problema se enfocaron en: el uso de herramientas de PLN como tokenización, lematización, *stemming*, *stopwords* y N-gramas, para el preprocesamiento de los documentos; el uso de la librería *gensim* utilizada en el entrenamiento y procesamiento del modelo; y finalmente el uso librerías de visualización como *pyLDAvis, matplotlib y wordcloud,* para mejorar la interpretación de los resultados.

En los resultados tenemos que: la selección del número de tópicos igual a 21 tiene el mejor balance entre interpretabilidad y descubrimiento de conocimiento; los tópicos generados permiten ser descritos, de tal manera que se puede tener un entendimiento global de estado actual de a investigación del tema, y además que se puede hacer una revisión descriptiva, en la que es posible evaluar el comportamiento a través del tiempo de los tópicos y cuáles son los principales *Journals* que publican alrededor del tema de interés.

Finalmente, los resultados obtenidos muestran que, el algoritmo LDA es adecuado para la solución de problemas de clasificación de texto, después de validar el resultado obtenido con una muestra aleatoria estratificada en la que se obtuvo un 69.35% de asertividad, para la correcta clasificación del modelo de manera global.

1. INTRODUCCIÓN

1.1 Contexto y Antecedentes

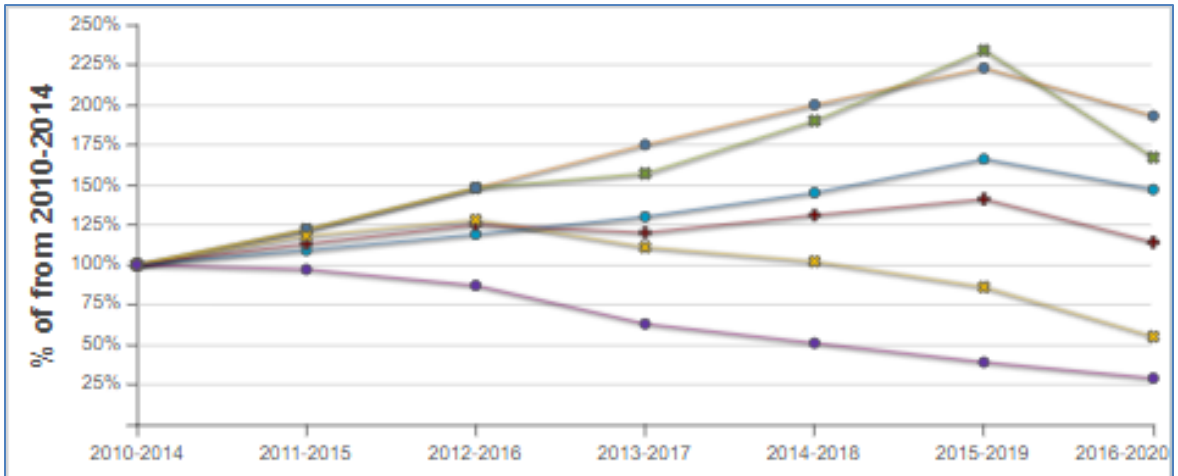
Los desarrollos de ciencia y tecnología son producto de la necesidad de lograr resolver los problemas y desafíos a los que se enfrenta día a día la sociedad. “El gasto en ciencia, tecnología e innovación; se asocia significativamente con el crecimiento de la productividad total de los factores e impacta positivamente en el crecimiento de un país” (Edquist & Henrekson, 2017). En consecuencia, se puede observar cómo los países con economías desarrolladas destinan un mayor porcentaje de su PIB en Investigación, Desarrollo e Innovación (I+D+i), en comparación a países con economías menos desarrolladas, logrando progresos científicos y tecnológicos de gran impacto, y estableciendo constantemente nuevos límites de desarrollo, manteniendo así el interés en la inversión en (I+D+i).

En Colombia, desde el año 1991 con la creación de Colciencias, y especialmente a partir del año 2009 con la promulgación de la Ley 1286 que eleva a Colciencias a Departamento Administrativo encargado de promover las políticas públicas para fomentar la ciencia, tecnología e innovación, la investigación comienza a tener un interés mayor de cara a la globalización, competitividad y visibilidad internacional de investigadores e investigación del país; lo anterior fue fortalecido con la entrada en funcionamiento en el año 2020 del Ministerio de Ciencia y Tecnología, cuya tarea es buscar y favorecer la productividad, la competitividad y el emprendimiento, con el fin de contribuir al desarrollo y crecimiento del país a través de la (I+D+i).

En el contexto actual y de acuerdo a datos publicados por la revista Dinero, Colombia invierte cerca del 0.7% del PIB en materia de innovación, del cual aproximadamente el 35% proviene del sector privado (Dinero, 2018), y aunque esto se encuentra lejos de los valores por encima del 2% de los países que más invierten

en (I+D+i) según la Organización para la Cooperación y el Desarrollo Económicos (OCED), en el ámbito de la investigación se evidencia un crecimiento primero en la producción científica indexada en revistas de alto impacto, pasando de 16.181 documentos indexados en el periodo 2010-2014 a 26.881 en el periodo 2015-2019 y segundo en la calidad en el contenido de los *Papers* publicados pasando de 231 *Top Papers* a 514 para los mismos periodos de referencia según la herramienta analítica para investigación *InCites* (ver Figura 1-1.), mejorando la reputación, escalafón y visibilidad tanto de investigadores como de Instituciones del país, las cuales a través de sus procesos de investigación constante generan nuevo conocimiento, o mejoran el conocimiento científico existente.

Uno de los factores más importantes al momento de empezar una investigación, es realizar una exploración sobre los diferentes medios existentes para publicar un producto académico. Lo anterior se debe de realizar, teniendo en cuenta la temática producto de la investigación y los términos y condiciones del medio en el que se quiere publicar. En esta etapa de la investigación es común que los investigadores utilicen herramientas como *Journal citation report (JCR), Scimago journal report (SJR), Wizdom y Dimensions*, las cuales permiten tener un acercamiento global del estado de pertinencia y actualidad de la investigación. Estas herramientas proporcionan métricas que clasifican la producción de conocimiento más influyente de los diferentes países del mundo, siguen el desarrollo de los principales macro temas y asignan un factor de clasificación cuya intención es generar un ranking en cuanto a la calidad de la producción académica, y es aquí donde cobra importancia la indexación en revistas de alto impacto y alta valoración académica, cuyo objetivo final es trasmitir la reputación y credibilidad de la revista en la que se desea publicar al producto de investigación que se va a encontrar publicado en esta.



Normalized	2010-2014	2011-2015	2012-2016	2013-2017	2014-2018	2015-2019	2016-2020
Total Papers	100%	109%	119%	130%	145%	166%	147%
Total Citations	100%	122%	148%	157%	190%	234%	167%
Total Citations per Paper	100%	113%	125%	120%	131%	141%	114%
Top Papers	100%	122%	148%	175%	200%	223%	193%
Citations to Top	100%	118%	128%	111%	102%	86%	55%
Citations per Top	100%	97%	87%	63%	51%	39%	29%

Raw	2010-2014	2011-2015	2012-2016	2013-2017	2014-2018	2015-2019	2016-2020
Total Papers	16,181	17,574	19,184	21,093	23,428	26,881	23,709
Total Citations	87,291	106,870	129,548	137,032	165,822	204,591	145,872
Total Citations per Paper	5.39	6.06	6.75	6.50	7.08	7.61	6.15
Top Papers	231	281	341	405	462	514	445
Citations to Top	94,600	111,803	121,195	104,979	96,602	81,659	52,378
Citations per Top	409.52	397.88	355.41	259.21	209.10	158.87	117.70

Figura 1-1 *Comportamiento de las publicaciones académicas recuperadas por InCites para Colombia desde el año 2010 hasta el primer semestre de 2020*

En Colombia existen criterios mínimos que determinan las características de una revista científica, incluyendo la gestión editorial, la validación del proceso de evaluación, la visibilidad e impacto, los cuales se explican a profundidad por Colciencias a través de la plataforma Publindex; a nivel latinoamericano existen índices similares como Scielo y Latindex los cuales cumplen con la misma misión, tomando como base un nivel geográfico más amplio; así mismo, existen grandes editores como Elsevier que sirven como publicadores a través de bases de datos como ScienceDirect y agregadores como es el caso de EBSCO y ProQuest, que agrupan contenidos de diferentes revistas de alto impacto a nivel científico, bajo altos estándares de publicación propios o de cada editorial agregada, lo que garantiza un nivel de calidad adecuado y llamativo para los investigadores e instituciones.

Otro de los factores claves a tener en cuenta en el proceso de investigación, es la recopilación de las referencias sobre las cuales se van a generar las bases de la propuesta a desarrollar. Durante este proceso, es común que los investigadores realicen una búsqueda intensiva de información relacionada a su tema de interés en bases de datos especializadas que han sido desarrolladas para este propósito. En estas y dependiendo del área, el investigador puede encontrar una cantidad importante de información, la cual con el avance del internet y el desarrollo de nuevas tecnologías y servicios para el acceso a la información, crece a niveles exponenciales como lo mencionan (Westgate et al., 2015) y lo reafirman (Greenville et al., 2017), generando un proceso de recopilación largo y cuya depuración se realiza de manera individual, inspeccionando cada uno de los artículos que han sido descargados y determinando si este artículo es o no de relevancia para su propuesta de investigación.

Teniendo en cuenta lo mencionado anteriormente, el proceso exploratorio que permite identificar el estado actual del tema de investigación de interés, las principales revistas científicas que lo abordan, y las colaboraciones que actualmente

se llevan a cabo, se convierte en una capacidad fundamental para el proceso investigativo y el desarrollo científico. Lo que genera diversas necesidades dentro de las que se encuentra la sintetización de la información, de tal manera que se puedan identificar tópicos de investigación con mayor facilidad y el estado actual de estos en la literatura.

Un adelanto bastante útil, ha sido utilizar métodos de clasificación automática de textos, provenientes del área del aprendizaje automático. Algunos de estos métodos, permiten descubrir esquemas y tendencias en colecciones de documentos (*Corpus*), en los que se identifican combinaciones de palabras y se ayuda a interpretar las ideas clave discutidas dentro de un Corpus. En definitiva, la capacidad de estas herramientas para generar ideas conceptuales disponibles tradicionalmente solo a través de la revisión esquemática de los textos, es de gran importancia, lo que permite visualizar una oportunidad desde el área del *Machine Learning*, desarrollando una herramienta de apoyo al proceso de síntesis y clasificación de información académica que sirva para facilitar la identificación de textos relevantes y posibles tópicos de investigación que se encuentren embebidos en la literatura.

1.2 Planteamiento del Problema

Una de las principales actividades que se realizan al iniciar un proceso de investigación, es la revisión del estado actual del tema de estudio en la literatura, se revisa si el tópico de interés ha sido ampliamente estudiado, es un tema emergente o si es una propuesta completamente nueva, con el fin de identificar oportunidades y direcciones de investigación futura, logrando así un entendimiento de cómo el área de interés ha cambiado con el tiempo.

Debido a las altas tasas de publicación científica (Westgate et al., 2015) (Greenville et al., 2017), la revisión previa del estado del arte de una investigación se vuelve una actividad que demanda cada vez más tiempo y esfuerzo por parte de los

investigadores, por lo que es necesario que esta revisión se realice de manera más rápida y sin perder la profundidad. Ante esta situación es importante proveer a los investigadores de herramientas orientadas a generar productos de investigación con mayor impacto, asistiendo en la síntesis de información y permitiendo identificar posibles oportunidades de investigación.

1.3 Objetivo General

Desarrollar un modelo de aprendizaje no supervisado para determinar temas de investigación, con el fin de apoyar el proceso de investigación durante la fase inicial de exploración.

1.4 Objetivos Específicos

1. Plantear el protocolo de adquisición del corpus de documentos a analizar.
2. Aplicar técnicas de aprendizaje no supervisado para la identificación de tópicos de interés en la base de textos analizada.
3. Evaluar la calidad y pertinencia del modelo desarrollado en la determinación de temas de investigación relevantes.

2. ANTECEDENTES

2.1 Marco Teórico

2.1.1 Resúmenes de textos y análisis de tópicos

Con la llegada de la “Era de la Información”, caracterizada por el uso del internet y las redes sociales, han aparecido nuevas formas para compartir y consumir información. Esto se ha convertido en un valioso *input* para las empresas en su necesidad de caracterizar clientes y lograr descubrir requerimientos de manera rápida y dinámica, y en el ámbito académico, ha permitido crear y reforzar redes entre investigadores y romper barreras de acceso a la información científica.

Lo anterior nos encamina hacia el concepto de “*text summarization*” (resumen de texto), que es un concepto extremadamente importante en el análisis de texto, el cual es usado en el contexto de la analítica supliendo la necesidad del acercamiento inmediato a un documento, lo que resulta de gran valor en diferentes sectores pues se logra acortar y resumir documentos de gran tamaño conservando el tema clave del documento. Esta información es presentada a consumidores y clientes para que puedan entender en cortos momentos de tiempo el tema principal de un determinado documento, lo cual es análogo a un “*Elevator Pitch*”, donde se debe proporcionar un resumen rápido que describa un proceso, producto, servicio o negocio, asegurando que conserve los temas y valores importantes centrales.

Dentro de las técnicas principales para el resumen y la extracción usado métodos de aprendizaje no supervisados nos encontramos las siguientes *Keyphrase Extraction*, *Topic Modeling* y *Automated Document Summarization* (Sarkar, 2019). Cada una de estas técnicas se utiliza en un contexto de problema diferente como lo mencionaremos a continuación en este documento.

2.1.1.1 Extracción de palabras clave

Esta es una técnica dentro del ámbito de la compresión del texto, usada para extraer información de documentos no estructurados, sobre la que se han propuesto una cantidad de enfoques efectivos y se han logrado buenos resultados..(Wen et al., 2014). La extracción de frase clave (*Keyphrase Extraction)*, conocida también como extracción de terminología (*Terminology Extraction)*, es el proceso de extraer términos o frases clave de un cuerpo de texto no estructurado para capturar sus temas centrales.

El resultado de este proceso son palabras o grupos de palabras generadas a partir de la construcción de n-gramas cuya clasificación se basa en la función de frecuencia de aparición de estos. Otros enfoques como los de (Barker & Cornacchia, 2000) y (Witten et al., 1999), nos permiten obtener una clasificación de N frases clave en función de sus pesos TF-IDF (*Term Frequency – Inverse Document Frequency*) a través un proceso de dos fases. Primero, se realiza la extracción de todos sintagmas nominales (*Noun phrase*) utilizando análisis sintáctico superficial (*Shallow Parsing*); y segundo, se calculan las ponderaciones TF-IDF para cada fragmento y recuperar las frases ponderadas más altas.

2.1.1.2 Resumen automático de textos

El resumen automático de textos (*Automated Document Summarization*) implica una serie de algoritmos capaces de crear resúmenes válidos teniendo en cuenta variables como la longitud, el estilo o la sintaxis, cuyo principal objetivo es encontrar un subconjunto representativo del texto original, presentado a modo de resumen ejecutivo en un solo documento, de tal manera que se contenga la idea central del documento original desde los puntos de vista semántico y conceptual. Existen dos enfoques principales en materia de resumen automático de textos, la extracción y la abstracción.

Las técnicas basadas en extracción utilizan conceptos matemáticos y estadísticos como *Singular Value Decomposition* (SVD) para extraer palabras u oraciones del texto original con el fin de crear el resumen. El resultado es un resumen con texto extraído del documento original, de ahí que se toma el nombre de la técnica basada en la extracción.

Las técnicas basadas en abstracción son más recientes, y utilizan técnicas de generación de lenguaje natural (NLG), junto con técnicas de aprendizaje profundo (*Deep Learning*), para crear un resumen por si misma similar a lo que un humano escribiría.

2.1.1.3 Análisis de tópicos

El análisis de tópicos (*Topic Model*) es un modelo probabilístico que contiene información sobre tópicos en un texto. Un "tópico" consiste en un grupo de palabras que frecuentemente ocurren juntas y que comparten el mismo tema ("Advanced Analytical Theory and Methods," 2015). Lo más importante es que el modelado de tópicos crea categorías temáticas sin necesidad de definir temas a priori. Es decir, que a diferencia de los sistemas de clasificación bibliográficos tradicionales, el modelado de temas determina la lista completa de temas a través del análisis de las ocurrencias de palabras a lo largo de un corpus de textos.

El modelado de tópicos implica la extracción de características de términos de documentos y el uso de estructuras y marcos matemáticos como la factorización matricial y el *Singular Value Decomposition* (SVD) para generar clústeres o grupos de términos que se pueden distinguir entre sí y formar tópicos. Estos tópicos se pueden utilizar para interpretar los temas principales de un texto y establecer conexiones semánticas entre palabras que se presentan con frecuencia en varios documentos. Existen varios marcos y algoritmos para construir modelos de temas. Los más comunes son *Latent Semantic Indexing* (Indexación semántica latente), *Latent Dirichlet Allocation* (Asignación de Dirichlet latente), *Non-negative matrix*

factorization (Factorización de matriz no negativa), *Hierarchical Dirichlet process* (proceso de Dirichlet jerárquica) y *Dynamic topic models (*Modelo dinámico de temas*)* (Srinivasa-Desikan, 2018)

En este documento se utilizará LDA (*Latent Dirichlet Allocation*) como método para acercarse a la solución del problema de identificación de tópicos, dado que “es el modelo de tópicos más Utilizado” (Korshunova et al., 2019), permite una buena interpretación de los datos y existe amplia disponibilidad de literatura referente al tema, lo que para este caso permite hacer comparaciones metodológicas relacionadas al tratamiento de este tipo de problemas.

2.1.2 LDA (Latent Dirichlet Allocation)

Latent Dirichlet Allocation es un modelo probabilístico generativo de tópicos. propuesto por (Blei et al., 2003) “Este modelo asume que los documentos se representan como mezclas aleatorias sobre tópicos latentes, en donde cada tópico se representa por una distribución de probabilidades sobre un vocabulario fijo de palabras” (Hernández et al., 2015). La distribución de tópicos para cada documento es generada a partir de una distribución de Dirichlet, permitiendo que un documento pertenezca en menor o mayor grado a varios tópicos, cada uno con un peso diferente.

LDA presenta 3 niveles de jerarquías Palabras, Documentos y Corpus (colección de documentos). Para entender mejor esto (Castillo, 2019), muestra un esquema didáctico el cual se presenta a continuación:

- *Documento 1: Tuve un sándwich de mantequilla de maní para el desayuno.*
- *Documento 2: Me gusta comer almendras, cacahuetes y nueces.*
- *Documento 3: Mi vecino consiguió un pequeño perro ayer.*
- *Documento 4: Los gatos y los perros son enemigos mortales.*
- *Documento 5: No debes alimentar con cacahuetes a tu perro.*

Los algoritmos de los modelos de tópicos analizan las palabras de una colección de documentos para descubrir los temas que en estos se trata, y cómo estos están conectados entre sí. Así, en el corpus de documentos anterior, LDA descubrirá 2 temas al analizar la distribución de las palabras dentro de cada documento:

- *Tópico 1: 30% maní, 15% almendras, 10% desayuno …*
- *Tópico 2: 20% perro, 10% gato, 5% cacahuete …*

A su vez se tiene que los temas se encuentran distribuidos de la siguiente manera en los documentos:

- *Documentos 1 y 2: 100% tópico 1.*
- *Documentos 3 y 4: 100% tópico 2.*
- *Documento 5: 70% Tópico 1, 30 % tópico 2.*

De esta manera los documentos son clasificados en tópicos y se puede proceder con la interpretación de estos. Adicionalmente se debe de realizar el análisis correspondiente en cuanto a los documentos que se agrupan en cada tópico y a la interpretación de cada uno de los tópicos identificados.

En la Figura 2-1. se ilustra el resultado del ejemplo planteado.

Figura 2-1. *Representación gráfica de la distribución de los tópicos en cada documento (Castillo, 2019)*

2.1.2.1 Representación del modelo LDA

A continuación, se mencionaras las notaciones y definiciones del modelo (ver Figura 2-2.):

- V es el número de palabras del vocabulario.
- w(d,n) es la enésima palabra del documento d, w(d,n) es un número entero en {1, ..., V} que indexa todas las palabras posibles.
- z(d,n): es el tópico de la enésima palabra en el documento d, z(d,n) es un número entero en {1, ..., K} que indica el tópico de la enésima palabra en el documento d.
- α: es el parámetro Dirichlet de la distribución Θ.
- β: es el parámetro Dirichlet de la distribución φ.
- φ(K): es la distribución de palabras del tópico K, φ(K) ~ Dir (β) con el vector de parámetros β > 0.
- Θ(d): distribución de tópicos del documento d, Θ(d) ~ Dir (α) donde Dir (α) es la distribución de Dirichlet con el vector de parámetros α > 0.
- φ y Θ se distribuyen Dirichlet; z y w se distribuyen multinomial.
- Hay N x M diferentes variables que representan las palabras observadas en los diferentes documentos.
- Hay K tópicos totales (la cantidad K está definida previamente).
- Hay M documentos totales.
- Hay N palabras en cada documento.
- En este modelo únicas variables observadas son las palabras en los documentos.

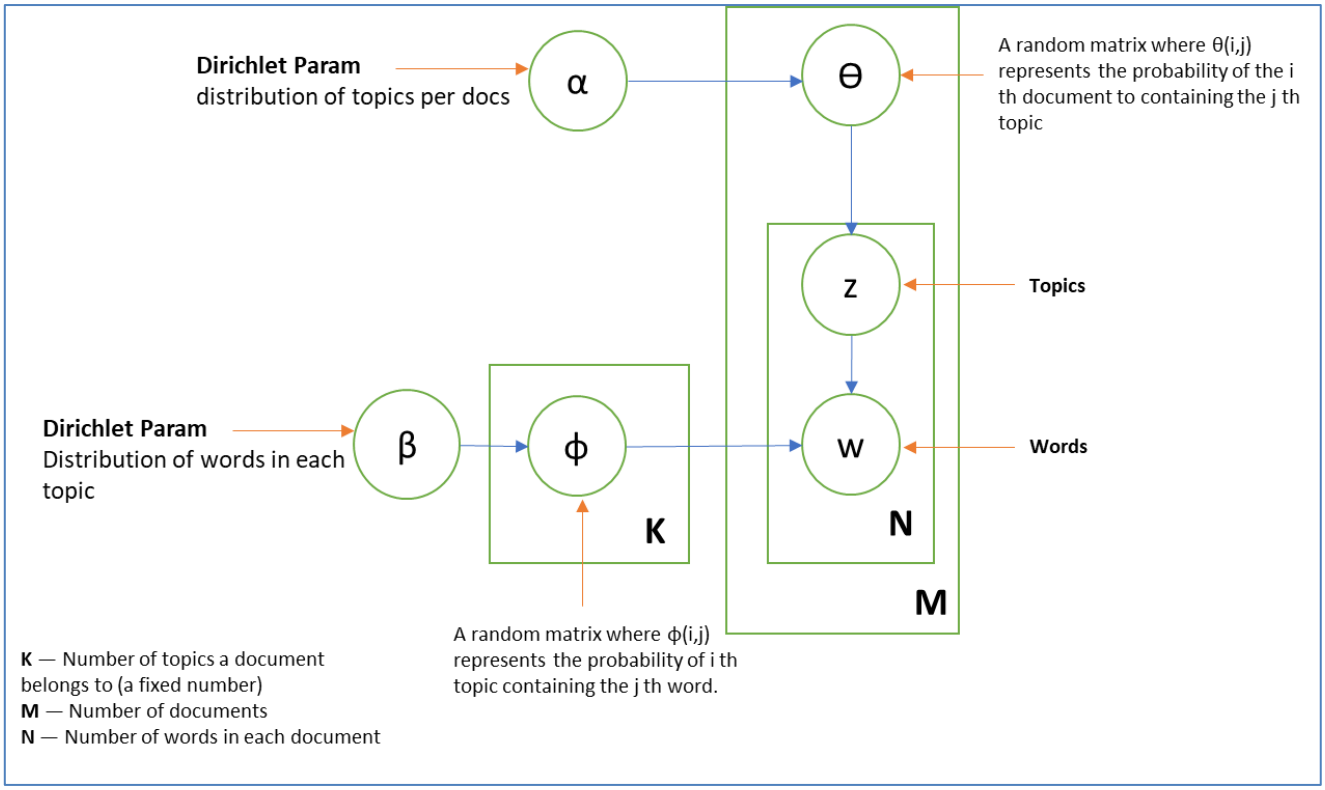


Figura 2-2. *Representación gráfica del modelo LDA (Blei et al., 2003).*

- Cada tópico es una distribución sobre palabras (ver Figura 2-3.).
- Cada documento es una mezcla de tópicos (ver Figura 2-3.).
- El tópico para cada palabra, la distribución por tópico para cada documento y la distribución de palabras por tópicos son variables latentes (variables que no se observan directamente, sino que son inferidas) en este modelo.

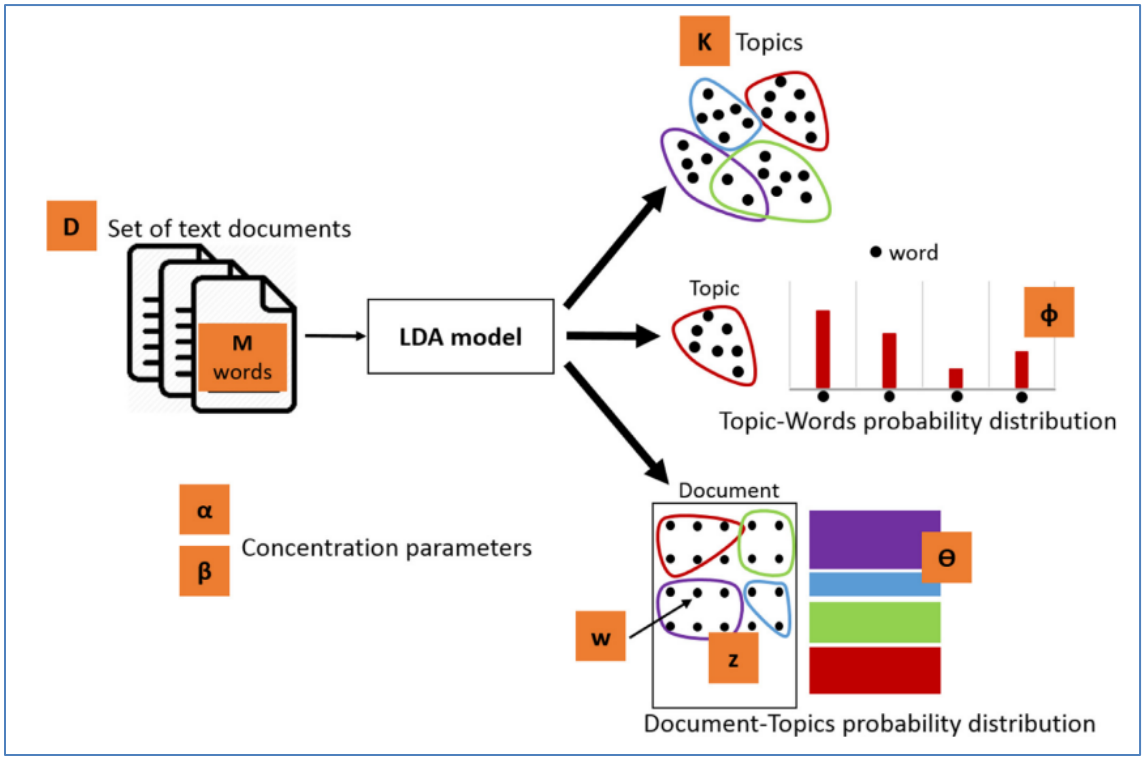


Figura 2-3. *Diagrama de interpretación de tópicos y documentos del modelo LDA (Nabli et al., 2018)*

Para un mejor entendimiento del modelo se pueden tratar φ y Θ como matrices creadas a partir de la descomposición de los documentos originales. En esta mirada Θ, se compone de filas determinadas por documentos y columnas determinadas por tópicos, así mismo, φ se compone de filas determinadas por tópicos y columnas determinadas por palabras. Por lo tanto, $\varphi_1, \ldots, \varphi_K$ se hace referencia a un grupo de filas, cada una de las cuales representa una distribución sobre palabras, e igualmente $\Theta_1, \ldots, \Theta_M$ se refiere a un grupo de filas, cada una de las cuales representa una distribución sobre tópicos.

2.1.2.2 Proceso Generativo

Al inicio de esta sesión se introdujo un ejemplo sencillo sobre la idea básica detrás del modelo LDA, ahora vamos a analizar de qué manera se concibe en el modelo que los documentos son generados. En LDA los documentos se representan como

mezclas aleatorias sobre tópicos latentes, donde cada tópico se define por una distribución sobre todas las palabras. Este método asume el proceso generativo para un corpus D que consiste en M documentos cada uno de longitud N_d (Blei et al., 2003).

A continuación, se detallan los pasos para generar un documento (Hui, 2019):

1. Para cada tópico,

(a) Definir una distribución sobre las palabras φ(K)~Dir (β).

Figura 2-4. *Definir una distribución sobre las palabras*

2. Para cada documento,

(a) Definir un vector de proporciones de tópicos Θ(d)~Dir (α).

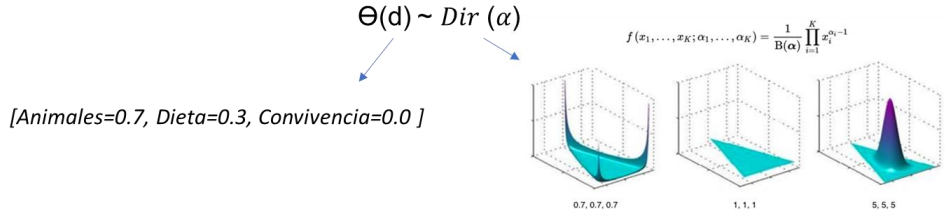


Figura 2-5. *Definir un vector de proporciones de tópicos Θ(d)~Dir (α).*

(b) Para cada palabra en el documento generado,

(i) Se elige aleatoriamente un tópico desde la distribución de tópicos. Para ello se utiliza una distribución multinomial $z_{(d,n)}$~Mult (Θ_d).

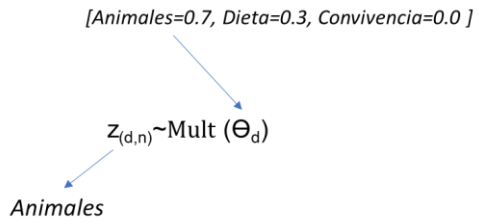


Figura 2-6. *Elección aleatoria de un tópico desde la distribución de tópicos.*

(ii) Según el tópico elegido, se elige aleatoriamente una palabra de la correspondiente distribución en el diccionario $w_{(d,n)}$~Mult ($\varphi_{z(d,n)}$).

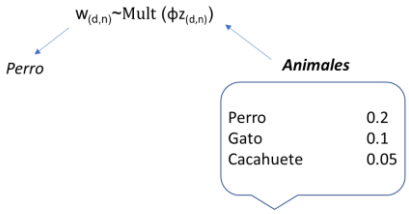


Figura 2-7. *Elección aleatoria de una palabra de la correspondiente distribución en el diccionario*

3. Este proceso se repite tantas veces como palabras tenga el documento generado.

2.1.2.3 Distribución de Dirichlet

La distribución de Dirichlet, es una distribución de probabilidad que puede considerarse como una generalización multivariada de la distribución Beta (Lin, 2016), la cual describe $k \geq 2$ Variables $X_1, \dots X_k$, en donde cada $x_i \in (0,1)$, $\sum_{i=1}^{N} x_i = 1$ y esta parametrizada por un vector de valores positivos $\alpha = (\alpha_1, \dots, \alpha_K)$ en donde α puede tomar valores decimales. En la Figura 2-8, propuesta por (Ganegedara, 2018), se puede observar cómo la distribución de θ cambia con valores diferentes de α. Mayores valores de α ubican los datos en la mitad del triángulo, mientras que menores valores de α ubican los datos en las esquinas.

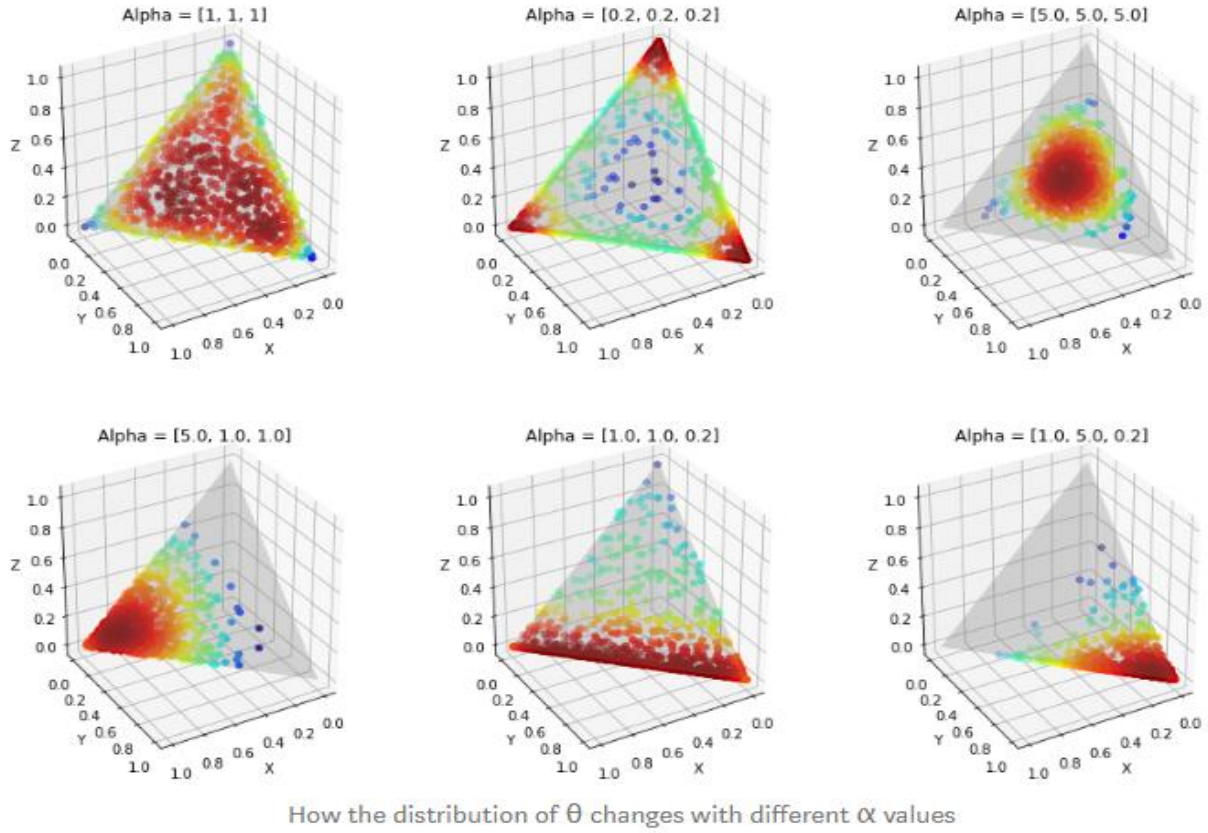


Figura 2-8. *Como la distribución de θ cambia con valores diferentes de α (Ganegedara, 2018)*.

2.2 Estado del arte.

2.2.1 Criterios de Búsqueda

Habiendo definido que la herramienta sobre la que se va a trabajar en este documento es LDA (*Latent Dirichlet Allocation*), y con el fin de hacer una evaluación del uso reciente y de marcos que se basen en este acercamiento, se procedió a buscar en las bases de datos académicas, documentos relacionados a este tema y que tuvieran aplicación a la clasificación de elementos académicos, Adicionalmente, Teniendo en cuenta que LDA es un modelo que lleva vigente desde el 2003, el periodo de publicación de artículos se filtró a publicaciones de entre 2004 y 2019.

Sobre los resultados obtenidos, se definieron las siguientes dimensiones con el fin de clasificar y comparar los elementos más relevantes encontrados (ver Tabla1).:

- Origen de los datos
- Documentos Evaluados
- Técnica usada
- Comparación con otros métodos
- Áreas de estudio
- Problema

2.2.2 Trabajos relacionados

(Griffiths & Steyvers, 2004) desarrollan un método en el que se utiliza información de *Proceedings of the National Academy of Sciences of the United States of America*, (PNAS), para ilustrar las relaciones entre diferentes disciplinas científicas, evaluando tendencias y "temas candentes" mediante el análisis de la dinámica de los temas (análisis de tópicos), y el uso de asignaciones de palabras a los tópicos para resaltar el contenido semántico de los documentos. Para lo anterior se propone LDA como método, destacando su conveniencia de uso dada la capacidad del modelo para proporcionar medidas cuantitativas, las cuales se pueden usar para identificar el contenido en los documentos, rastrear cambios en el contenido a lo largo del tiempo y expresar la similitud entre estos.

Se concluye, que este método permite descubrir aspectos significativos de la estructura de la ciencia, revelando la existencia de relaciones entre artículos científicos en diferentes disciplinas y facilitando la comprensión de la información contenida en grandes colecciones de documentos, que incluyen, registros de correo electrónico, grupos de noticias y en general documentos contenidos en internet.

(Yau et al., 2014) realizan una comparación entre algoritmos no supervisados tales como LDA y sus extensiones. Con este propósito, se evalúan diferentes colecciones de documentos académicos relacionados y no relacionados, y se valora la

capacidad del método de modelado de tópicos para distinguir estructuras del texto, utilizando el *recall* y la precisión como métricas de evaluación.

Los autores concluyen que los algoritmos de análisis de tópicos pueden clasificar documentos con una precisión significativa. Sin embargo, es necesario llevar a cabo más investigaciones centradas tanto en algoritmos como en procesos para validar el modelado de temas como una herramienta práctica para la investigación cienciométrica.

(Westgate et al., 2015) presentan LDA como un método de búsqueda y procesamiento de texto que puede ser usado para investigar tendencias e identificar vacíos potenciales de la investigación en la literatura científica. En este sentido, concluyen que los métodos de análisis de texto necesitan ser usados con cuidado, pero también que pueden proporcionar información que es complementaria a aquella dada por las revisiones sistemáticas y los metaanálisis, lo que incrementa la capacidad de los científicos para sintetizar la investigación.

(Li et al., 2019) presentan el desarrollo de un Framework en el que utiliza documentos científicos y patentes como recursos de entrada e integra técnicas de análisis de tópicos para pronosticar tendencias en tecnológicas y las distancias que existe entre la producción científica y el desarrollo de tecnologías. Se menciona el análisis de tópicos, especialmente LDA como herramienta que juega un papel importante en la exploración de la tendencia de desarrollo de la tecnología. Y se menciona que estas técnicas han sido usadas por algunos académicos para descubrir los temas técnicos que están implícitos en los documentos para estudiar las tendencias tecnológicas.

(Luo, 2019) propone un método de análisis de sentimientos de texto que combina la representación de texto de Asignación de Dirichlet Latente (LDA) y las redes neuronales convolucionales (CNN), con el fin de mejorar el rendimiento del análisis de sentimientos públicos de Internet. Primero, los textos de revisión se recopilan de

la red para el pre-procesamiento. Luego, se usa el modelo de tópicos LDA para obtener el vector de características de texto corto (distribución de tópicos). Y finalmente se usa un modelo CNN recurrente (con una base GRU – *Gated Recurrent Unit*) como clasificador. De acuerdo con la matriz de características de entrada, el GRU-CNN fortalece la relación entre palabras y palabras, texto y texto, para lograr una clasificación de texto de alta precisión. Los resultados de la simulación muestran que este método puede mejorar efectivamente la precisión de la clasificación del sentimiento del texto.

2.2.1. Tabla de resumen de criterios.

Tabla 1 *Resumen de criterios*

	(Griffiths & Steyvers)	**(Yau et all)**	**(Westgate et al)**	**(Li et al)**	**(Luo)**	**Propuesta**
Año	2004	2014	2015	2019	2019	2020
Origen de los datos	PNAS	WOS	No menciona	Derwent Innovations Index (DII) + WOS	Encuesta abierta sobre Notebook en Jingdong shopping mall micro-blog de análisis de sentimientos	Bases de datos académicas
Documentos Evaluados	Abstracts	Artículos científicos	Abstracts, + Artículos científicos	Patentes + Artículos científicos	Cometarios + archivos XML	Artículos Científicos
Técnica usada	LDA	LDA, Correlated topic models (CTM), Hierarchical Latent Dirichlet Allocation (Hierarchical LDA) and the Hierarchical Dirichlet Process (HDP)	LDA + (analysis of topic similarity, generality, popularity, and research gaps)	Lingo algorithm	LDA + CNN con unidad recurrente (GRU)	LDA + (identificación de principales revistas, comportamientos por año y similitud entre documentos)
Comparación con otros métodos	NO	SI	NO	NO	SI	NO
Áreas de estudio	Multi-área	Multi-área	Ecología	Perovskite solar cell technology	Mercadeo	Tecnologías en educación

Problema	Clasificación	Clasificación	Identificación de Tendencias	Identificación de Tendencias	Clasificación	Clasificación + Identificación de tendencias

Como se puede observar en esta revisión, el modelo LDA ha sido ampliamente utilizado en temas de clasificación del conocimiento, la literatura en este sentido es amplia y permite tener referentes para la validación metodológica y de resultados, también se puede observar que se presentan como desafíos nuevos acercamientos que incluyen CNN para mejorar la precisión de las clasificaciones por lo que el modelo aún tiene grandes posibilidades en el desarrollo de la investigación.

3. METODOLOGÍA

3.1 Modelo de ciclo de vida incremental

En el desarrollo de este trabajo se utilizará el modelo de ciclo de vida incremental. Este modelo, permite ir añadiendo sucesivamente funcionalidad al producto por medio de etapas o iteraciones en un lapso predeterminado. Al final de cada iteración, se habrá completado un entregable y el producto final será la acumulación de funcionalidades construida en las iteraciones.

Algunas de las bondades del modelo se mencionan a continuación:

- Visión de avance.
- Aprendizaje en cada iteración.
- Flexibilidad.
- En caso de ser requerido se puede retomar a partir de una iteración anterior.

Las fases del modelo incremental son:

- Comunicación: inicio del proyecto y recopilación de requisitos.
- Planeación: estimación del itinerario y los seguimientos.

- Modelado: análisis y diseño.
- Construcción: código y se prueba.
- Despliegue: Entrega, soporte y retroalimentación.
- Evaluación del modelo.

Teniendo en cuenta los tiempos disponibles, se proponen 4 incrementos con una ventana de tiempo de 1 mes por incremento.

3.1.1 Incremento 1

En este incremento se pretende determinar cuáles serán las fuentes de información óptimas para el desarrollo de la tarea teniendo en cuenta las restricciones que cada fuente de información pueda llegar a tener. Lo anterior se logra mediante la exploración de la información que se pueda recuperar en las búsquedas.

Apunta al desarrollo del Objetivo Especifico 1.

3.1.2 Incremento 2

En este incremento se considera realizar un proceso de evaluación de la herramienta y posibles modelos alternos que sean aplicables para la solución del tipo de tarea que se plantea en este trabajo. Esto se pretende realizar a través de una exploración bibliográfica

Apunta al desarrollo del Objetivo Especifico 2.

3.1.3 Incremento 3

Habiendo cumplido con los incrementos anteriores, en este incremento se debe de tener claro el mejor modelo a utilizar con el que se va a construir la herramienta.

Apunta al desarrollo del Objetivo Especifico 2

3.1.4 Incremento 4

Una vez la herramienta empiece a generar resultados se realizarán validaciones con diferentes expertos temáticos quienes determinaran la facilidad de interpretación de los resultados y la utilidad de la herramienta.

Apunta al desarrollo del Objetivo Especifico 3.

4. RECOLECCIÓN DE DATOS

La recolección de datos es un elemento fundamental en este proyecto, de la cantidad de información que se logre recuperar y del tiempo que esto requiera depende en gran medida que esta herramienta sea o no atractiva a la hora de que un investigador desee realizar su revisión bibliográfica previa a su investigación. En este momento del proyecto, se quiere ejemplificar de la manera más exacta posible, la realidad en el proceso de recolección de información que un investigador llevaría a cabo, con la intención de tener un punto de comparación real frente al método tradicional.

4.1 Selección del área de conocimiento a evaluar

Dado el amplio universo de conocimiento existente, es necesario acotar el área del saber que se desea evaluar, con el fin de que el algoritmo logre realizar agrupaciones temáticas que sean realmente útiles desde lo práctico y que la información que se logre obtener pueda ser evaluada por expertos temáticos (en nuestro caso expertos temáticos en educación mediadas por Tics), quienes determinaran finalmente la pertinencia del algoritmo.

Para seleccionar el tema que va a acompañar este trabajo, se decidió que éste se va a alinear con una necesidad puntual de descubrir tópicos relacionados con

"*Tecnología en la educación"* propuesto por la Universidad ICESI como área de interés en la búsqueda continua del fortalecimiento de su modelo educativo. Lo anterior teniendo en consideración la disponibilidad de expertos temáticos, la inmersión del autor en el entorno educativo y la información disponible para realizar este trabajo.

4.2 Características y selección de fuentes de Información

La fuente por defecto para la revisión del estado actual del conocimiento es el artículo científico. En especial aquellos que se encuentran indexados en revistas de alto impacto tal como se menciona al inicio de este documento. El procedimiento más común es revisar en herramientas como WOS o Scopus® cuáles son los artículos y autores más citados, y qué artículos tienen mayor coincidencia con los términos de búsqueda del área de interés. Estas bases de datos de información académica tienen la característica de ser fuentes ampliamente aceptadas por la comunidad científica para recopilar la información de fuentes cuyo nivel de exhaustividad en la revisión y publicación es muy fuerte, así como por manejar imparcialidad respecto a la información que en estas se encuentra.

4.3 Estrategia de recolección de datos

Continuando con el propósito de aterrizar este ejercicio a la realidad, es importante mencionar que dentro del proceso de búsqueda de recursos de información es crucial que el investigador tenga ampliamente desarrolladas, las competencias informacionales que le permitan realizar búsquedas en bases de datos académicas. Esto incluye: la construcción de ecuaciones de búsqueda, las cuales se componen de elementos de interés temático, expresados en forma de palabras claves y de uno o más operadores booleanos (*AND, OR, NOT),* que permiten conectar de forma lógica conceptos o grupos de términos para así ampliar, limitar o definir las búsquedas; el entendimiento básico sobre filtros y la descarga y gestión de metadatos asociados a las búsquedas.

Para este proyecto se revisaron sugerencias de expertos y bibliografía relacionada con el tema propuesto (ver Tabla 2), de los primeros se adoptó la sugerencia de utilizar la aproximación pedagógica denominada "*Computer-supported collaborative learning"* (CSCL) como marco de referencia para el planteamiento de las ecuaciones de búsqueda; y de la revisión bibliográfica se logró extraer palabras claves relacionadas para enriquecer las ecuaciones.

Tabla 2 *Recolección de palabras claves en la bibliografía relacionada*

Titulo	Autor	Año	Términos Destacados
Digital Technologies: Sustainable Innovations for Improving Teaching and Learning (Sampson et al., 2018)	Demetrios Sampson	2018	3d-model, abilities, abstraction, algorithm, audio files, audio-recordings, automation, BCI controlled, blended-learning, charts, coding, cognitive, communication, competences, computational thinking, computer, creativity, curricula, data analysis, data collection, data flow, data representation, databases, digital library systems, e-book, e-readers search, file formats, files video, flipped classroom, gaming, gathering information, graphs, images, information literacy, interactive, internet, inverted classroom, iterative thinking, laptops, learning environment, learning management system, lectures, libraries, LMS, logic, media, mobile learning, mobile tablets, multimedia, open education, parallel thinking, parallelization, PDF, pedagogical learn technologies, playing, programming, reading, recursive thinking, robotics, simulation, simulations, skills, smartphones, social software tools, STEM, students, tablet, teachers, text, video, video game, video-based learning platforms, virtual reality (VR), visual, visual representation, visualizations, WEB, words
Emergent Practices and Material Conditions in Learning and Teaching with Technologies (Cerratto Pargman & Jahnke, 2019)	Teresa Cerratto Pargman	2019	acquisition, apps, artefacts, augmented reality, camera, classroom, cognition, collaborative learning, communication, competences, competences searching, computers, creative critique, curriculum, digital, digitization, educational, educational values, field trips, film recorder, game-based, inquiry based learning, internet, knowledge, learning spaces, learning teaching, mobile devices, museums, outdoor learning, projector, research, tablets, teacher, technologies, video chat, virtual, white-board
Learning and Teaching with Technology in the Knowledge Society New Literacy, Collaboration and Digital Content (Iinuma, 2016a)	Mizuho Iinuma	2016	3D content, 3D models, 3D visualize, active learning, algorithm, assessing credibility of information, audio, audio contents, audio creation, blogs, bulletin boards, chats, classroom, co-editing, collaboration, collaboration classroom, collaborative, collaborative skills, communication, competences, computer collaborative, computer educator, computer literacy, cooperation, co-writing, creation, creativity, critical thinking, CSCL activities, curriculum, database, deep thinking, digital art, digital content, digital publishing, discussion, education, email, excel sheet, feedback giving, finding, flipped classroom, flipped learning, google earth, graphic design, group management skills, groupware, ICT, images, information assessing, information literacy, inquiry skills, instant messages, institutions, internet, interpersonal skills, IT access, IT literacy, knowledge society, learning, lifelong learners, linking, literacies,

			literacy education, media literacies, media literacy, mobile, mooc, moving, moving images, multimedia, multimedia interpretation, multimedia production, multimodal literacy, navigating, negotiation, notebook, open education, participatory media, power point, questioning, reading, reading skills, role-playing, search engines, searching, sharing, skill, SNS, social media websites, social network services, social networking, software, student, teaching, technology, text, time management, video blogs, video conferencing, video content, video contents, video creation, video editing, video lectures, video streaming, visual, WEB, webpage, wiki, wikis, word processing, writing, WWW, youtube
New Digital Technology in Education (Ng, 2015)	Wan Ng	2015	3D printing, analytics, BYOD, BYOT, classroom, cloud computing, computer, content creation, curriculum, data gathering, data mining, digital content, digital literacy, digital resources, digital society, digital tools, education, educator, profesional development, flipped classroom, games, gamification, knowledge, learn, mobile learning, MOOC, pedagogy, skills, social media, social networking internet, student, teach, technology, digital, online learning, WEB
Shaping the University of the Future (Marshall, 2018)	Stephen James Marshall	2018	active learning, agile universities, agility, alumn cost, audio, automation, BYOD, celular, cloud, communication tools, communicative, curriculum, devices, digital native, economy, e-learning, email, expensive, experiences, feedback systems, flexible, global university, google scholar, higher education, images, independent learning, individual courses, information, information literacy, information resource, information seeking, information skills, innovation, institution, intercultural, internet, invention, journal subscriptions, learn, lectures, LMS, media, mobile, modern universities, modular, MOOC, multimodal, networking pedagogical, online collaboration tools, online education, online knowledge, online sources, open education, open online courses, open source, open technology, personal network, price, print, remote, remote monitoring, research, robotics, search, smart glasses, social media, social networks, society, software, spaces rented, stakeholders, students, subscriptions, surgeons, tablet, teach, technology, technology skills, telephones, tele-presence, texting, tools, transformation, university, video, video lectures, virtual university, virtually, visual, WEB, weekend sessions, wireless, word processing

El resultado de este ejercicio, fue una variación de la propuesta de (Iinuma, 2016b) de la definición de “ambiente típico CSCL” (ver Figura 4-1.). en esta variación se incluyen, el lugar en donde se realiza la actividad y se adicionan tecnologías adicionales de interés para este proyecto (ver Tabla 3). Debido a que la gran mayoría de la información que se encuentra en WoS y Scopus® está en idioma inglés, las búsquedas y en adelante todo el ejercicio se realizará utilizando este idioma.

	Collaborative skills	IT Access	Literacies
CSCL activities	Interpersonal skills • Discussion • Negotiation • Communication	Internet Search Engines Google Earth Power Point Excel Sheet Groupware Wiki, SNS Email, chats Instant messages	Reading • Navigating • Searching • Finding • Information assessing
	Inquiry skills • Feedback giving • Questioning		
	Group management skills • Time management • Role-playing		Writing • Co-writing • Co-editing • Digital publishing

Figura 4-1. *(Collaborative skill, IT access, and literacies in CSCL) Esquema propuesto por (Iinuma, 2016b)*

Tabla 3 *Esquema propuesto para el planteamiento de ecuaciones en la búsqueda de tópicos relacionados con "Technology in Education"*

	Collaborative Skills	**Technologies**	**Spaces**	**Literacies**
Technology In Education	Interpersonal Skills • Abstraction • Argumentation • Communication • Discusion • Feedback Giving • Flexible • Negotiation	Multimedia and Interactive Technology *3D, Audio, Augmented Reality, Automation, Digital, Games, Gamification, Google Earth, Graphic Design, Graphs, Images, Interactive, Media, Multimedia, Robotics, Simulation, Technology, Video, Virtual Reality (VR), Virtually, Visual, Camera*	Physical Spaces • Blended Learning • Classroom • Curriculum • Education • Flipped Classroom • Global University • Inverted Classroom • Learning Spaces • Libraries • Modern Universities • Museums • Outdoor • Spaces Rented	Reading • Computer Literacy • CSCL Activities • Digital Literacy • Finding • Information Assessing • Information Literacy • Information Resource • Information Seeking • IT Literacy • Media Literacy • Navigating • Online Knowledge
	Inquiry Skills • Active Learning • Assessing Credibility Of Information • Computational Thinking • Creation • Creativity	Coding and Data Analysis *Algorithm, Analytics, Apps, Cloud, Software, Coding, Data, Feedback Systems, Programming, Open Source*		

	• Critical Thinking • Data Gathering • Deep Thinking • Gathering Information • Independent Learning • Innovation • Inquiry • Invention • Learn • Lifelong Learners • Online Learning • Questioning • Research • Syntthesize • Transformation	Social Media and Office Tools *Blogs, Youtube, Word Processing, WWW, Social Media, Social Networks WEB, Webpage, PDF, Power Point, Excel Sheet, Email, Wiki, Chats, Ebook, Instant Messages, Office Tools, Search Engines, Text*	• University Virtual Places • 3D Content • Digital Resources • E-Learning • Google Scholar • ICT Education • Individual Courses • IT Access • Journal Subscriptions • LMS • Mobile Learning • MOOC • Multimodal • Online Education • Open Education • Open Online Courses • Remote • Remote Monitoring • Tele-Presence • Virtual University • Weekend Sessions	• Online Sources • Searching Writing • Co-Editing • Co-Writing • Digital Publishing • Word Processing
	Group Work Skills • Agility • Co-Editing • Collaboration • Collaborative • Cooperation • Co-Writing • Digital Society • Group Management • Intercultural • Personal Network • Role-Playing • Social Networking • Stakeholders • Teach • Time Management	Hardware and Smart Technology *Internet, Smart Glasses, Smartphones, White-Board, Telephones, Tablet, BCI Controlled, BYOD, BYOT, Cellular, Charts, Computer, Communication Tools, Wireless, Artefacts, Groupware, Mobile, Open Technology, Print*		

Con el esquema planteado, se puede visualizar una estructura conceptual sobre la que se van a realizar las ecuaciones de búsqueda. Cabe anotar que este esquema es una guía y en ningún momento pretende acotar o definir específicamente el orden de los criterios o la cantidad de estos en las ecuaciones.

4.4 Selección del origen de la información y definición de las ecuaciones de búsqueda

La cantidad de información que se puede recopilar y la facilidad de acceso a los metadatos de esta son parámetros determinantes en el momento de seleccionar la fuente desde la que se van a buscar y posteriormente descargar los datos. Durante la exploración de los distintos proveedores de búsqueda de artículos académicos se determinó que Scopus® es el proveedor que tiene las mejores características en cuanto a la relación que hay entre la cantidad de artículos indexados y las opciones de descarga de los metadatos, la cual supera con creces a sus competidores como se puede evidenciar en la Tabla 4.

Tabla 4 *Comparación entre herramientas de investigación cienciométrica*

	Dimensions	**WoS**	**Scopus®**
Tipo de acceso	Acceso libre	Suscripción institucional	Suscripción Institucional
Cantidad máxima de metadatos en una sola descarga	500	500	2000, 20.000 a través de solicitud por correo

Definición de las ecuaciones de búsqueda

Posterior a haber seleccionado la herramienta sobre la que se va a realizar el proceso de búsqueda de la información, se procede a estructurar las ecuaciones con base en la exploración realizada el punto 4.3, para esto se construyó una

bitácora (Ver Tabla 5) en donde se almacenan los datos, que el autor considera más representativos durante las consultas realizadas.

Durante el proceso de definición de las ecuaciones, se realizaron algunos ajustes tales como:

- Exclusión de palabras *(AND NOT school AND NOT press AND NOT immediat* and not children)*. Se excluyeron las palabras school, immediat* (el asterisco representa las posibles variaciones después de la letra "t"), y children. Lo anterior debido a que se quiere recuperar información referente a estudios superiores y que la frase "media" recuperaba muchos resultados con la frase immediate que no es una frase relevante en este estudio; también se excluyó la palara "press" debido a que existen *journals* que tienen como nombre alguna de las palabras que hacen parte de la ecuación de búsqueda, continuado con de la frase "press" p.ej. (education press).
- Filtro por año y tipo de recurso *(LIMIT-TO (PUBYEAR, 2021) OR LIMIT-TO (PUBYEAR, 2020) OR LIMIT-TO (PUBYEAR, 2019) OR LIMIT-TO (PUBYEAR, 2018) OR LIMIT-TO (PUBYEAR, 2017) OR LIMIT-TO (PUBYEAR, 2016) OR LIMIT-TO (PUBYEAR, 2015)) AND (LIMIT-TO (DOCTYPE,"ar"))*. Se limitó la búsqueda a los años 2015-2021 y adicionalmente se filtró la búsqueda para que solo recuperara recursos de tipo "Articulo", Esto debido que el objetivo de este estudio es recuperar los tópicos más recientes en esta área y que la información recuperada en un artículo es más compleja que en otros recursos de información.

Como se puede observar en la Tabla 5, el número de artículos que recupera esta herramienta es de 15.144. Este número, es un numero de referencia, toda vez que las instituciones académicas, centros de investigación, entre otros, solo tienen acceso a una cantidad limitada de bases de datos, que según su presupuesto reduce significativamente la cantidad de descarga de artículos en texto completo.

Tabla 5 *Bitácora de Búsqueda de información*

Periodo de búsqueda	Ecuación	Área(s) Principales	Observaciones
Todos	TITLE-ABS-KEY((educat* or * "Blended Learning " or Classroom or "Flipped Classroom" or "Inverted Classroom" or "Learning Spaces" or Libraries or Outdoor or "Digital Resources" or E-Learning or "ICT Education" or "Mobile Learning" or MOOC or "Online Education" or "Open Education" or Remote or "Tele-Presence" or "Virtual University" or "Weekend Sessions") AND (technolo* or 3D or Audio or "Augmented Reality" or Automation or Digital or Games or Gamification or Google or Graph* or Image* or Interactive or *media or Robot* or Simulation or Video or "Virtual Reality" or VR or Virtually or Visual or Camera or Coding or Data or Algorithm or Analytics or Apps or Cloud or Software or Coding or Programming or "Social Media" or Blogs or Youtube or "Social Networks" or WEB or Webpage or Wiki or Ebook or "Instant Messages" or "Office Tools" or "Search Engines" or Internet or "Smart Glasses" or "Smartphones" or "White-Board" or Telephones or Tablet or "BCI Controlled" or BYOD or BYOT or Cellular or Computer or "Communication Tools" or Wireless or Artefacts or Groupware or Mobile or "Open Technology") AND (skills OR learn* OR teach* or Literacy) AND NOT school AND NOT press AND NOT immediat* and not children)	Social Sciences (37,101) Computer Science (34,688) Engineering (16,936) Mathematics (5,783) Medicine (5,081)	Cantidad de resultados 71.745 Fecha de Búsqueda 20/08/2020
(2015-2021)	TITLE-ABS-KEY((educat* or * "Blended Learning " or Classroom or "Flipped Classroom" or "Inverted Classroom" or "Learning Spaces" or Libraries or Outdoor or "Digital Resources" or E-Learning or "ICT Education" or "Mobile Learning" or MOOC or "Online Education" or "Open Education" or Remote or "Tele-Presence" or "Virtual University" or "Weekend Sessions") AND (technolo* or 3D or Audio or "Augmented Reality" or Automation or Digital or Games or Gamification or Google or Graph* or Image* or Interactive or *media or Robot* or Simulation or Video or "Virtual Reality" or VR or Virtually or Visual or Camera or Coding or Data or Algorithm or Analytics or Apps or Cloud or Software or Coding or Programming or "Social Media" or Blogs or Youtube or "Social Networks" or WEB or Webpage or Wiki or Ebook or "Instant Messages" or "Office Tools" or "Search Engines" or Internet or "Smart Glasses" or "Smartphones" or "White-Board" or Telephones or Tablet or "BCI Controlled" or BYOD or BYOT or Cellular or Computer or "Communication Tools" or Wireless or Artefacts or Groupware or Mobile or "Open Technology") AND (skills OR learn* OR teach* or Literacy) AND NOT school AND NOT press AND NOT immediat* and not children) AND (LIMIT-TO (PUBYEAR,2021) OR LIMIT-TO (PUBYEAR,2020) OR LIMIT-TO (PUBYEAR,2019) OR LIMIT-TO (PUBYEAR,2018) OR LIMIT-TO (PUBYEAR,2017) OR LIMIT-TO (PUBYEAR,2016) OR LIMIT-TO (PUBYEAR,2015)) AND (LIMIT-TO (DOCTYPE,"ar"))	Social Sciences (10,229) Computer Science (4,169) Engineering (2,518) Medicine (1,580) Arts and Humanities (1,326)	Cantidad de resultados 15.144 Fecha de Búsqueda 23/08/2020

4.5 Descarga y transformación de datos para la ingesta.

Tal como se mencionó anteriormente, la cantidad de información a la que se tiene acceso varía dependiendo de las suscripciones que el investigador tenga a su disposición. Por lo tanto, para recopilar la mayor cantidad de datos, en este documento se alimentó el algoritmo LDA de la siguiente manera:

Se realizó una descarga (ver Figura 4-2.) de los metadatos contenidos en la base de datos Scopus®, la cual entrega hasta 20.000 registros en formato *.csv*. Estos metadatos no son leídos de manera nativa por los diferentes gestores bibliográficos, por lo que fue necesario apoyarse en el programa Python para diseñar un algoritmo que permitiera transformar los archivos .csv a archivos .ris*,* los cuales sí son leídos por la mayoría de los gestores bibliográficos. En esta transformación fue de gran importancia lograr recuperar el DOI (identificador de objeto digital) pues con base en éste se logró enriquecer la información (el resumen y las palabras claves del autor para cada uno de los documentos) utilizando el programa Mendeley.

El objetivo de esta forma de recuperar la información es aprovechar las bondades que tiene el programa Mendeley en cuanto a la estructuración de la información, la cual queda prácticamente lista para ser ingresada al modelo LDA. Adicionalmente, de esta forma se va a evaluar la pertinencia del modelo para trabajar con la información contenida en el título, resumen y palabras claves del autor.

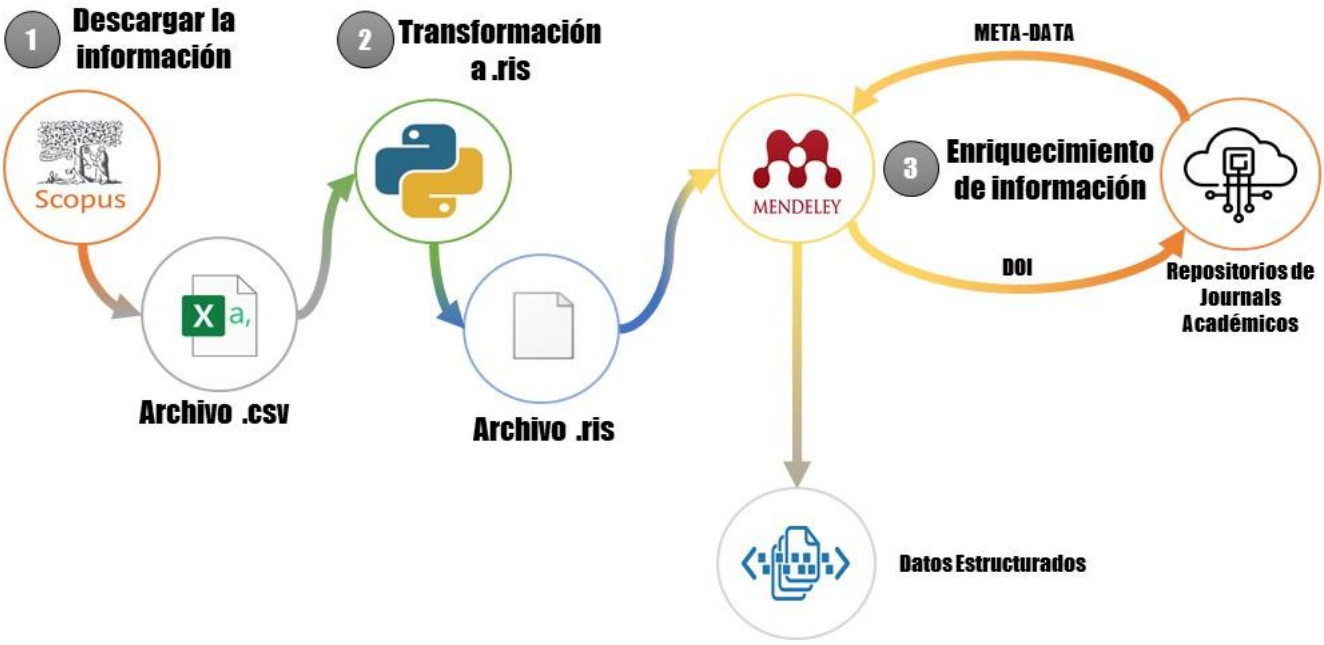


Figura 4-2. *Ruta para la preparación de los datos de entrada del algoritmo LDA desde los Metadatos recuperada por medio de la herramienta Scopus®*

4.6 Volumen de información recolectada.

A pesar de que los metadatos originales sufrieron una transformación y un posterior enriquecimiento, no hubo perdida en la información que se logró recolectar, tal como se muestra en la Tabla 6. Sin embargo, es posible que esta metodología de recolección de información tenga algunas perdidas, debido a que el identificador DOI (*Digital Object Identifier*) en ocasiones muy puntuales (sobre todo en artículos muy antiguos) tiene problemas con el retorno de los datos.

Tabla 6 *Volumen de información recolectada por cada metodología.*

No. de artículos Reportados desde Scopus®	Datos recolectados	Tipo de información	Observaciones
15.144	15.114	Metadatos	Título, palabras claves del autor y Resumen

5. MODELADO

5.1 Pre-procesamiento de los datos

previo al ingreso de los datos al modelo, es necesario que estos sean preparados para que el algoritmo LDA pueda interpretar de manera adecuada la información. Para esto, es necesario construir una matriz en la que cada documento se relacione con cada una de las palabras del corpus (unión de todos los documentos) y se realice un recuento de la cantidad de apariciones de cada palabra en cada documento.

Con el propósito de obtener mejores resultados, no todas las palabras del corpus son tenidas en cuenta ya que, al final del ejercicio, un menor número de palabras facilitará la interpretación y reducirá considerablemente el tamaño de la matriz, optimizando el tiempo de procesamiento. A continuación, se mencionarán algunas metodologías aplicadas para extraer palabras que generan ruido en el modelo.

5.1.1 Librería *Gensim*

Gensim es una libraría de Python usada en el modelado de tópicos (que incluye LDA y LSI), indexación de documentos, recuperación de similitudes y para trabajar con modelos vectoriales de palabras (como *Word2Vec* y *FastText*). una de sus principales ventajas, es que permite manejar archivos de texto grandes sin tener que cargar el archivo completo en la memoria, con lo que se puede manejar grandes colecciones de texto.

5.1.2 Tokenización

Esta tarea consiste en extraer de las oraciones las palabras, las cuales pasan a recibir el nombre de *token*. Durante este proceso se eliminan las puntuaciones y

caracteres innecesarios, un método dentro de la librería *gensim* útil para esto es *gensim.utils.simple_preprocess()*, el cual no solamente genera los *tokens*, sino que también los cambia en minúsculas y quita todo tipo de acentos.

En el desarrollo del presente trabajo, fue necesario utilizar expresiones regulares para adecuar los datos a las necesidades puntuales del ejercicio. Por ejemplo, fue necesario que la palabra e-learning, fuese cambiada por e_learning, pues si se hacia el proceso de tokenizar sin este cambio, la letra “e” desaparecía durante el proceso perdiendo el contexto particular de la palabra “e-learning”; igualmente, fue necesario eliminar los correos electrónicos de los autores, saltos de páginas y comillas.

5.1.3 *Stopwords*

En muchos idiomas existen palabras cuya función es conectar y dar sentido a las oraciones, pero que por sí solas no incorporan un significado. a estas palabras las vamos a definir como *Stopwords*. Comúnmente, las *Stopwords* suelen ser artículos, pronombres, preposiciones y conjunciones. Algunos ejemplos de estas palabras en el idioma ingles son *“a”, “the”, “is”, “are”, “and”,* etc.

En muchos problemas asociados al PLN (Procesamiento de lenguaje natural) como clasificación de texto o análisis de sentimientos, es bastante útil la eliminación de las *Stopwords*, pues no es deseado que en la construcción del modelo estas palabras ocupen espacio en nuestro diccionario, además pueden agregar ruido a los datos previniendo encontrar patrones inherentes. Realizar esta limpieza ayuda a optimizar el tiempo de procesamiento y a que la interpretación del modelo sea más “clara”.

En el caso de nuestro modelo adicionalmente a las *Stopwords,* que vienen predeterminadas en el método *stopwords.words('english'),* fue necesario agregar otras palabras tale como “http”, “https”, “www”, “doi”, “journal”, “magazine” y

“abstract”. lo anterior debido a que son palabras muy comunes en los artículos académicos.

5.1.4 Lematización y Stemming

5.1.4.1 Lematización

En la estructura de los diferentes idiomas, encontramos que hay palabras de diferente forma de escritura y que sin embargo representan el mismo sentido. En español, por ejemplo, sabemos que bailo, bailas, baila, bailamos, bailan son formas (conjugaciones) diferentes del mismo verbo (Bailar). Otro ejemplo es que gata, gato, gatita, gatos, gatitos y otros son formas diferentes de la palabra Gato. Por lo tanto, con el objetivo de mejorar el modelo, es necesario combinar todas estas variantes, en una sola palabra base. Esto es posible desarrollar, por medio de la lematización la cual, relaciona las variantes de las palabras con su forma base o lema.

La lematización es un proceso clave en muchas tareas de PNL, pero tiene dos desventajas principales. Primero, es un proceso que consume recursos (especialmente tiempo); en segundo lugar, suele ser probabilístico, por lo que en algunos casos obtendremos resultados inesperados. En el caso de este proyecto, se utilizó la biblioteca *Spacy* y su diccionario en inglés para realizar este proceso.

5.1.4.2 Raíz (Stemming)

Este proceso convierte palabras en raíces. Estas raíces son la parte invariable de las palabras relacionadas principalmente por su forma. En cierto modo se parece a la lematización, pero los resultados (las raíces) no tienen que ser palabras de un idioma. lo que para para las palabras cortas puede llegar a ser un problema en su interpretación. Por ejemplo, para las palabras tu cantas, ella canta, nosotros cantamos, la raíz seria cant.

El *Stemming* es significativamente más rápido desde el punto de vista del procesamiento que la lematización, tiene la ventaja de poder identificar las correspondencias entre palabras de diferentes tipos, por lo que es capaz de reconocer, que picante y picar tienen como raíz pic-. El *Stemming* es un método que permite reducir la cantidad de elementos que componen nuestros diccionarios, lo cual es extremadamente beneficioso para el modelo. En Python dentro de la librería *nltk* se encuentra el método *SnowballStemmer* que realiza esta tarea.

Una desventaja del *Stemming* es que al ser sus algoritmos más simples que los de lematización, pueden "cortar" demasiado la raíz y encontrar relaciones entre palabras que realmente no existen. También puede suceder que deje raíces demasiado extensas o específicas, y que tengamos más bien un déficit de raíces, en cuyo caso no se convierten palabras que deberían convertirse a la misma raíz (Fernández, 2019).

En esta fase de preprocesamiento, se decidió comparar las 2 alternativas (Lematización Vs *Stemm*) de reducción de dimensionalidad, el principal criterio fue la interpretabilidad de las palabras generadas. Para esto, se les pidió a 2 expertos temáticos, que revisaran las palabras que se generaron después de aplicar cada uno de los métodos y seleccionaran la alternativa que a juicio de ellos tuviera la mejor interpretabilidad, que para el caso de este trabajo fue Lematización (ver Tabla 7). Cabe anotar, que el hecho de que para este modelo la lematización genere resultados con mejor interpretación, ambos métodos son válidos durante el proceso de reducción de dimensionalidad y puede ocurrir que para otros *corpus* funcione mejor *Stemm*.

Tabla 7 *Matriz para la selección del mejor modelo.*

Variable	Experto	Método
Interpretabilidad	1	Lematización
Interpretabilidad	2	Lematización

5.1.5 N-gramas

Existen palabras, cuyo sentido se incrementa cuando esta se acompaña por una o más palabras complementarias, a esta asociación, en NLP se de denominan N-gramas, Los más comunes son los bigramas y los trigramas. Un ejemplo de esto son las frases como *"cloud_computing", "digital_process" y "higher_education"*. Al igual que en el punto anterior, la cantidad de N-gramas en este caso, fue evaluado con los expertos temáticos, quienes sugirieron manejar bigramas dada la interpretabilidad que se logró.

Una biblioteca bastante útil para construir los N-gramas es *gensim* cuyo método *models.Phrases*, permite seleccionar la cantidad de veces que se debe repetir un bigrama de todo el *corpus* para ser tomado en cuenta y definir un límite (*threshold*) cuya función es controlar la cantidad de frases que se forman. Para el caso de este modelo los valores que se seleccionaron fueron 10 para el primer parámetro, y 10 para el *threshold* cuyo valor se determinó a partir de la ejecución de diferentes valores, hasta que aparecieran términos que fueran interesantes desde el punto de vista temático.

5.2 LDA con *gensim* para análisis de metadatos

5.2.1 Parametrización

Después de haber pre-procesado la información, el modelo está listo para ser entrenado. Es importante mencionar que la solución al modelo LDA en *gensim*, está basado en el algoritmo de solución planteado por (Hoffman et al., 2010) *"Online learning for Latent Dirichlet Allocation"* el cual representa una alternativa al algoritmo de *Gibbs sampling*.

5.2.1.1 Cantidad de tópicos a elegir

Ya definiendo los parámetros, la primer gran incógnita que surge es ¿cuántos tópicos se van a generar?, para responder a esta pregunta, se debe de tener en cuenta, que aunque existen medidas como la perplejidad y la coherencia, las cuales consisten en medir qué tan bien una distribución de probabilidad o modelo de probabilidad predice una muestra y medir el grado de similitud semántica entre las palabras respectivamente; tal como lo menciona (Chang et al., 2009) valores como la perplejidad, en ocasiones no se encuentran correlacionados con el juicio humano.

En la determinación de la cantidad de tópicos a generar, es importante tener en cuenta, que no es necesario interpretar todos los tópicos generados. En este sentido, dependiendo del contexto, se puede buscar un equilibrio entre la interpretabilidad de los tópicos y la calidad del tópico. Así pues, la cantidad de tópicos generados puede aumentarse tanto como el problema requiera.

5.2.1.2 Selección del número de tópicos por medio de métricas

Como se mencionó anteriormente, desde el análisis numérico se propone un valor de referencia en la determinación del número óptimo de tópicos a seleccionar. Así pues, se evaluó el valor de la coherencia (*Coherence Value*) iterando sobre los parámetros *num_topics y decay*. Este último es un número entre (0.5, 1] que pondera qué porcentaje del valor lambda anterior se olvida cuando se itera cada nuevo documento. Corresponde a la variable Kappa del modelo de (Hoffman et al., 2010).

Para esta prueba, se utilizó la coherencia, como medida de referencia para identificar el número óptimo de tópicos. Así pues, se ensayaron valores desde 6 hasta 26, incrementando en 4 unidades cada vez para el numero de tópicos y los valores de decaimiento (*decay*) 0.5, 0.7 y 0.9. Los resultados muestran que los

mejores valores se logran con una decadencia de 0.5 y que a partir de 18 tópicos los valores de coherencia convergen como se evidencia en la figura 5-1.

Finalmente teniendo en cuenta que, desde el análisis numérico, aumentar el número de tópicos no alteraba significativamente el valor de coherencia (para el intervalo 18-26) y que según la teoría revisada hasta el momento aumentar el valor del número de tópicos de manera razonable nos puede definir mejor ciertos grupos de tópicos se decidió tomar como valores 18, 21 y 24 para el parámetro número de tópicos.

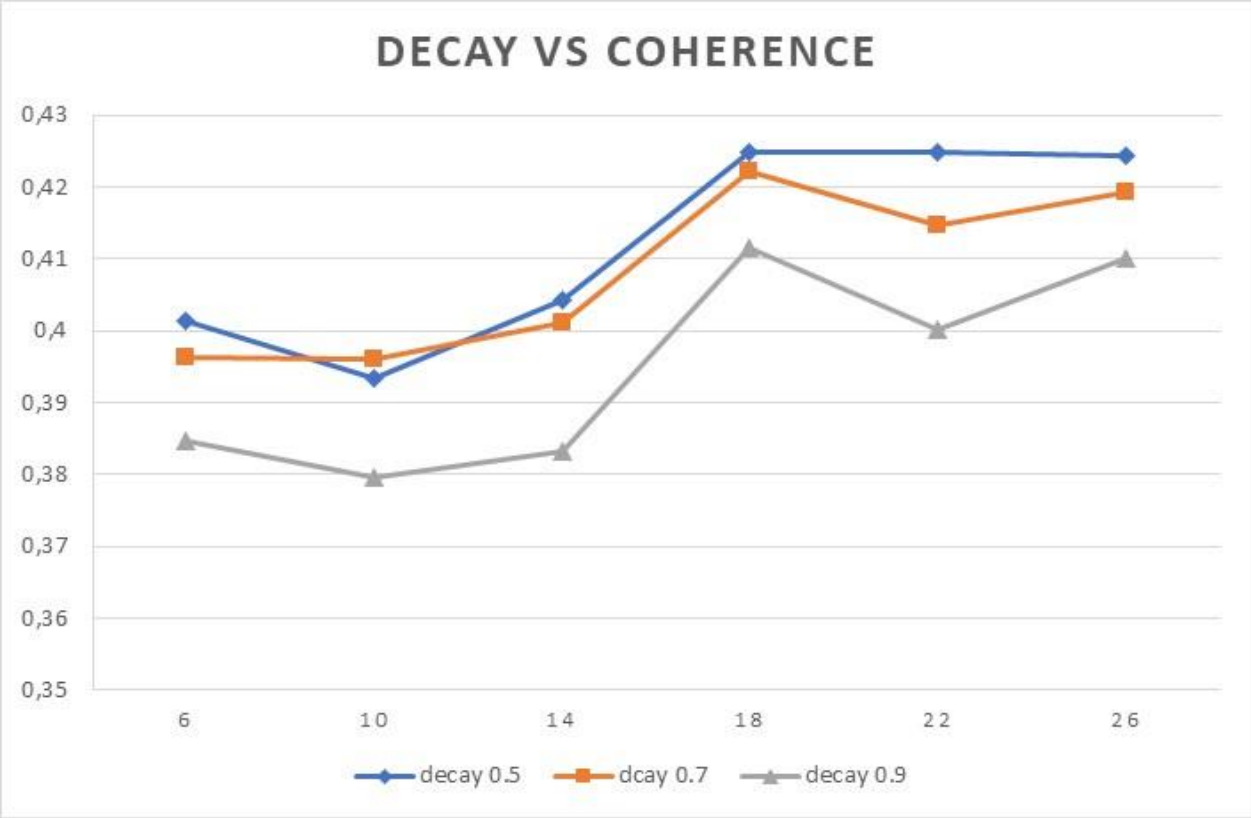


Figura 5-1 *determinación del número de tópicos a través de medidas de referencia*

5.2.1.3 Entrenamiento del Modelo

Habiendo definido la cantidad de tópicos y continuando con el supuesto de que las medidas de bondad del modelamiento nos permiten tener una aproximación de la calidad, se realiza el proceso de determinar los hiper-parámetros que nos permitan maximizar las medidas de bondad del modelo. En este caso el paquete *gensim* cuenta con dos parámetros en particular a considerar.

El primero, *passes,* análogo a las épocas, hace referencia al número de veces que se entrena el modelo en todo el corpus. En teoría una mayor cantidad de épocas permitiría que más documentos converjan. El segundo, *iterations,* controla la frecuencia con la que repetimos una ruta particular en cada documento durante el entrenamiento. Más técnicamente, controla cuántas iteraciones sin convergencia, se permite el ciclo del algoritmo de solución LDA (ver Figura 5-2.) que define los parámetros variacionales γ y φ, por lo que hay un límite sobre la duración del proceso si todos los documentos no han logrado la convergencia.

5.2.1.4 Métricas para revisar y resultado del entrenamiento

Una de las bondades de la librería *gensim*, es la capacidad de almacenar los datos de los procesos de entrenamiento en Logs (Archivo de registro). En este caso, fue posible obtener la información relacionada a la perplejidad, convergencia y coherencia, que el algoritmo LDA de *gensim* va generando a medida que se va entrenando. Lo anterior se logró a través de la construcción de un archivo en donde se fueron registrando estos eventos.

Algorithm 1: Variational Expectation-Maximization LDA

Input: Number of topics K
Corpus with M documents and N_d words in document d
Output: Model parameters: β, θ, z

initialize $\phi^0_{ni} := 1/k$ for all i in k and n in N_d
initialize $\gamma_i := \alpha_i + N/k$ for all i in k
initialize $\alpha := 50/k$
initialize $\beta_{ij} := 0$ for all i in k and j in V

//E-Step (determine ϕ and γ and compute expected likelihood)
loglikelihood := 0
for $d = 1$ **to** M
  repeat
    for $n = 1$ **to** N_d
      for $i = 1$ **to** K
        $\phi^{t+1}_{dni} := \beta_{iw_n} \exp\left(\Psi(\gamma^t_{di})\right)$
      endfor
      normalize ϕ^{t+1}_{dni} to sum to 1
    endfor
    $\gamma^{t+1} := \alpha + \sum_{n=1}^{N} \phi^{t+1}_{dn}$
  until convergence of ϕ_d and γ_d
  loglikelihood := loglikelihood + $L(\gamma, \phi; \alpha, \beta)$ // See equation BLAH
endfor

//M-Step (maximize the log likelihood of the variational distribution)
for $d = 1$ **to M**
  for $i = 1$ **to K**
    for $j = 1$ **to V**
      $\beta_{ij} := \phi_{dni} w_{dnj}$
    endfor
    normalize β_i to sum to 1
  endfor
endfor
estimate α via Eq. (8)

if loglikelihood converged **then**
  return parameters
else
  go back to E-step
endif

Figura 5-2. *Algoritmo de maximización esperada variacional LDA*

Al entrenar un modelo, es necesario definir las métricas más importantes y cuál es el rendimiento mínimo esperado. Como regla general, se debe de buscar el punto en el que las métricas relevantes se aplanan y usar esos puntos como valores predeterminados. Lo anterior, teniendo en cuenta que es poco probable que las métricas cambien significativamente sin que cambie la fuente de información o el preprocesamiento.

Con el objetivo de revisar el comportamiento del modelo, respecto a la variación de los valores de interés (*passes e iterations*) se alteraron los valores de estos parámetros así: 25 épocas (*passes)* e iteraciones (*iterations)* de 10, 20, 30, 50, 80, 100 y 150. De los resultados obtenidos se observó lo siguiente:

- La figura 5-3., muestra los cambios de la coherencia en las diferentes iteraciones de épocas seleccionadas, es posible notar que, en la primera época, la coherencia inicia en valores que van desde 0.428 hasta 0.445, lo que no denota una diferencia significativa. A partir de la segunda época, mejora significativamente y luego se estabiliza después de 5 épocas. Al final, después de 25 épocas, solo un valor (50 *iterations*) se ubica por debajo de 0.49, estando los demás valores entre 0.49 y 0.508.
- La figura 5-4., muestra los cambios de la perplejidad en las diferentes iteraciones de épocas seleccionadas, los valores para la primera época, se dividen en 2 grupos, el primero alrededor de 171.2 y el segundo alrededor de 182. A partir de la segunda época, los valores mejoran hasta y se aplana después de 10 iteraciones. Finalmente, los valores se agrupan alrededor de 157,3 con una desviación de 1.83.
- Los documentos que convergen, nos indica la cantidad de documentos que el algoritmo es capaz de clasificar correctamente por cada *batch* o grupo que se le pasa. En la Figura 5-5., se puede advertir, que los mejores valores para este parámetro de obtienen entre mayor cantidad de iteraciones se realicen,

mejorando significativamente a partir de 80 iteraciones. En este caso la curva también parece aplanarse a partir de 10 iteraciones.

Teniendo en cuenta los resultados obtenidos hasta el momento, y procurando mantener un buen compromiso entre rendimiento y el tiempo de entrenamiento se tomó como medida de referencia los siguientes valores *passes* = 10 e *iterations* = 80. Finalmente, y tal como la documentación de la librería *gensim* lo menciona, si se dispone de capacidad computacional y tiempo para entrenar el modelo, la cantidad de épocas e iteraciones puede ser tan grande como se quiera considerar. Sin embargo, realizar el análisis del comportamiento de las métricas es muy útil en cuanto a que ayuda a saber cuáles son los compromisos entre la capacidad y el tiempo de ejecución, lo que permite identificar una buena opción para el trabajo a realizar.

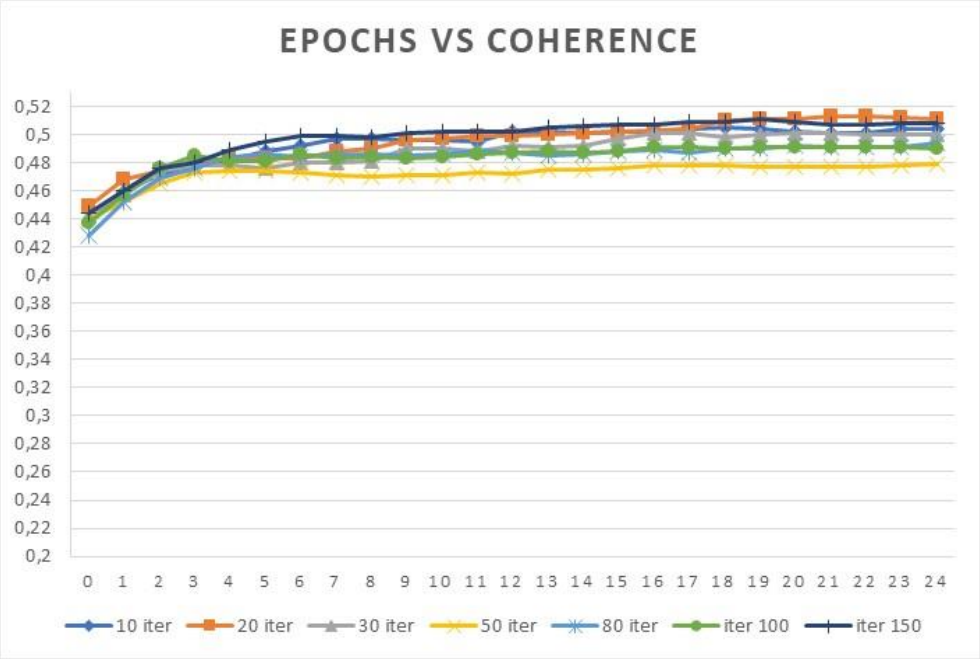


Figura 5-3 *Determinación del número de épocas e iteraciones del modelo.*

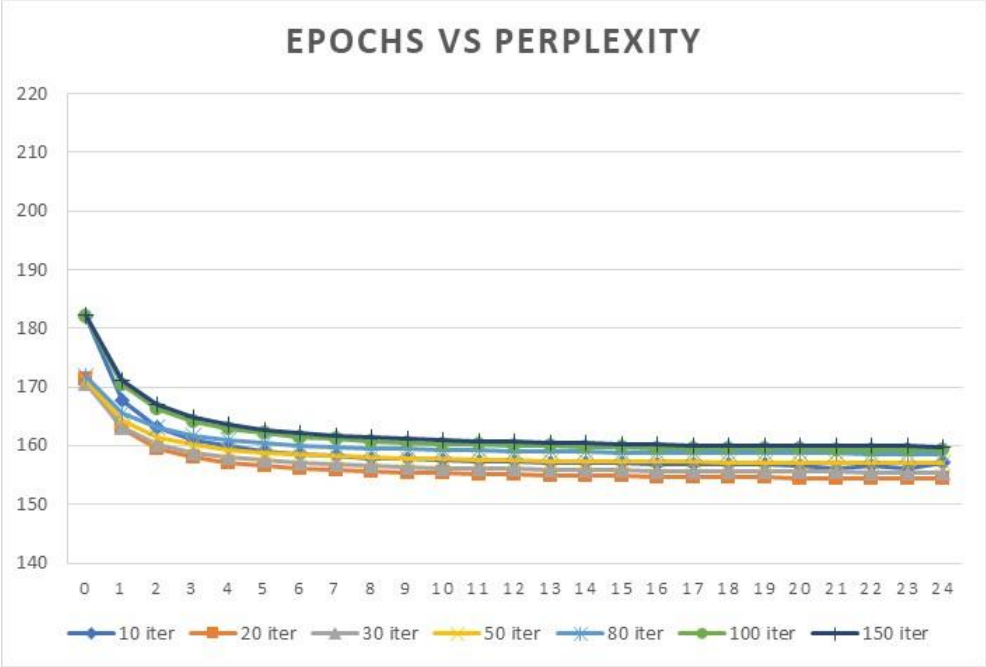


Figura 5-4 *Determinación del número de épocas e iteraciones del modelo.*

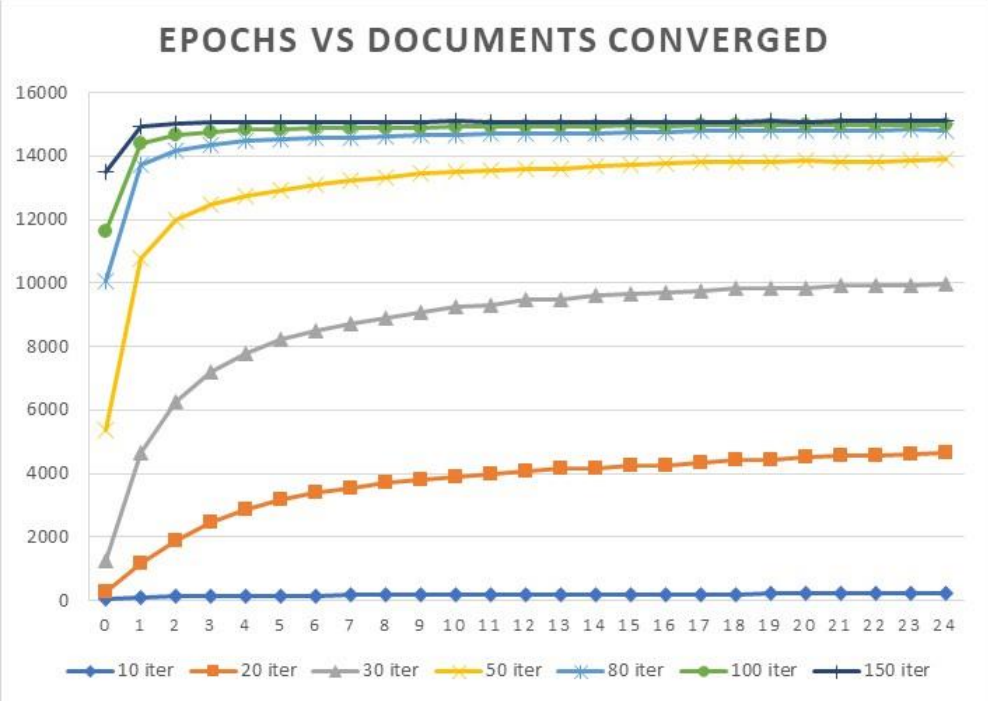


Figura 5-5 *Determinación del número de épocas e iteraciones del modelo.*

6. RESULTADOS

En este capítulo, se presentan los resultados de los 3 modelos de tópicos entrenados (18, 21 y 24 tópicos), y se determinará cuál de estos, según el criterio de los expertos temáticos, agrupa de mejor manera los tópicos. Sobre este modelo se realizarán análisis que permitan entender un poco más como se conforman los tópicos.

Para lograr este objetivo, se presentará para cada modelo a evaluar la distribución generada por la librería *pyLDAvis.* Está librería, está diseñada para ayudar a los usuarios a interpretar los temas en un modelo de tópicos que se ha ajustado a un corpus de datos de texto. El paquete extrae información de un modelo de temas de LDA entrenado, para generar una visualización interactiva (ver Figura 6-1.).

A continuación, se evaluarán los resultados generados por la librería *gensim* y la librería *pyLDAvis* con los expertos temáticos, quienes ayudarán a determinar con qué modelo continuaremos en el análisis. Después de esto, se va a identificar la agrupación de términos de cada tópico dado un porcentaje de composición, la cantidad de documentos por tópico, los tópicos de referencia para cada documento; adicionalmente, gracias a que los datos lo permiten, se va a relacionar algunos datos de la base de datos original, para evaluar, cómo se han comportado los tópicos en el tiempo y adicionalmente, desagregar esta información por *journals* académicos.

Finalmente, y como una herramienta adicional que permite identificar un documento nuevo dentro del corpus ya entrenado, se incluye la opción de asignar un documento a un tópico, con la cual es posible identificar dentro del universo existente qué autores y *journals* manejan temas similares y qué cantidad de documentos existen es esta sub-agrupación.

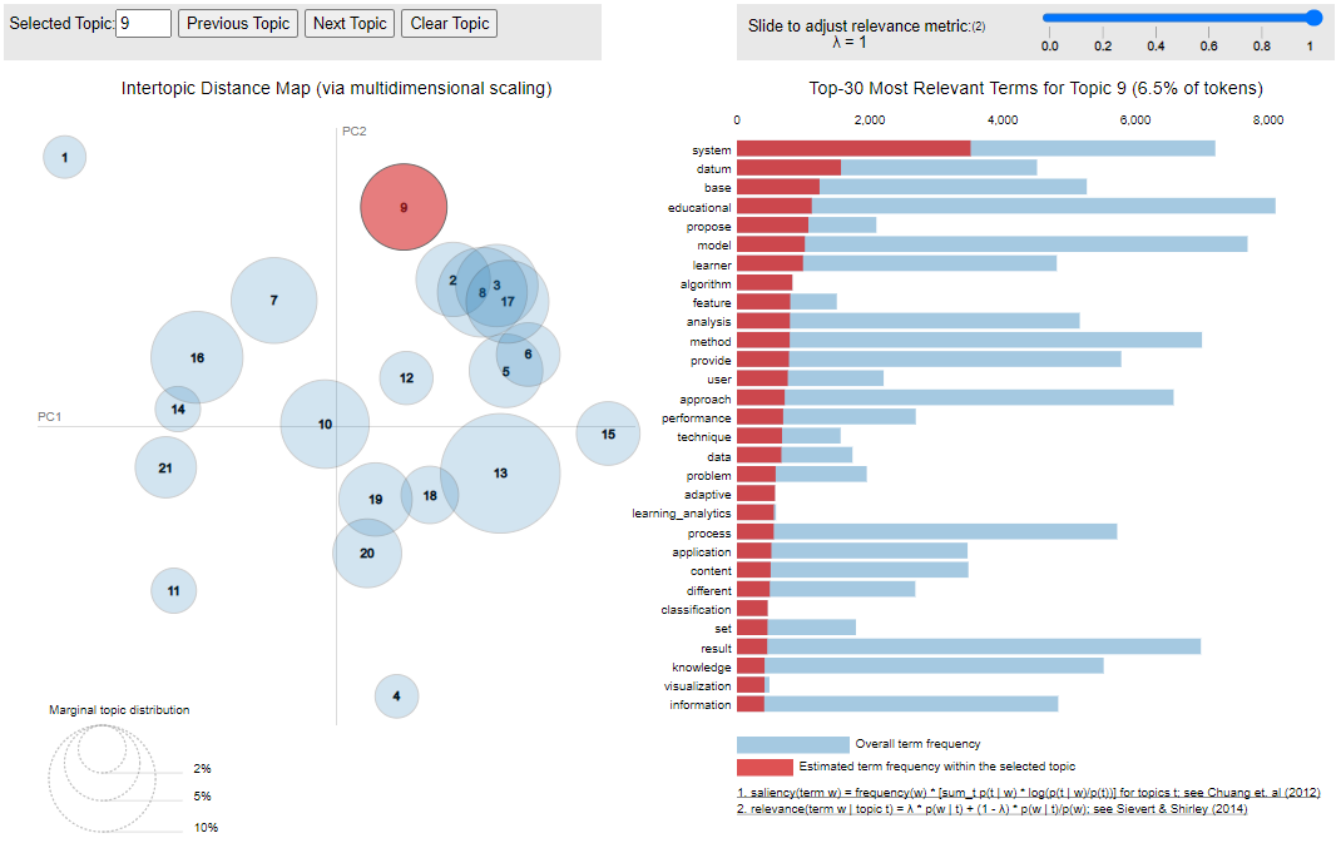


Figura 6-1. *visualización interactiva del paquete pyLDAvis con λ=1*

(cada burbuja representa un tópico el cual es descrito por las palabras que se encuentra a la derecha del gráfico)

6.1 Generalidades de la librería *pyLDAvis*

Debido a la importancia que tiene la visualización en este trabajo como recurso para interpretar la información del modelo, se presentan las principales generalidades a tener en cuenta a la hora de trabajar con la herramienta *pyLDAvis*. Esta herramienta proporciona dos piezas importantes de información. Primero, en el lado Izquierdo (ver Figura 6-1.), hace uso de la reducción de dimensionalidad, con la cual, a través de las distancias entre los círculos, se determina la similitud entre tópicos, adicionalmente, el área de cada uno de los círculos representa la distribución marginal del tópico, y depende del porcentaje del corpus que se clasifica en el tópico; Segundo, como se puede observar en el lado derecho de la Figura 6-1. se

presenta una barra deslizante la cual, presenta una puntuación de relevancia λ, la cual cuando es igual a 1, contiene los términos de mayor frecuencia por cada tópico respecto a todo el corpus, y a medida que este valor se disminuye y se acerca a 0 se identifican los términos que pertenecen de manera más exclusiva al tema, lo que permite poder interpretar los tópicos de una mejor manera (ver Figura 6-2.).

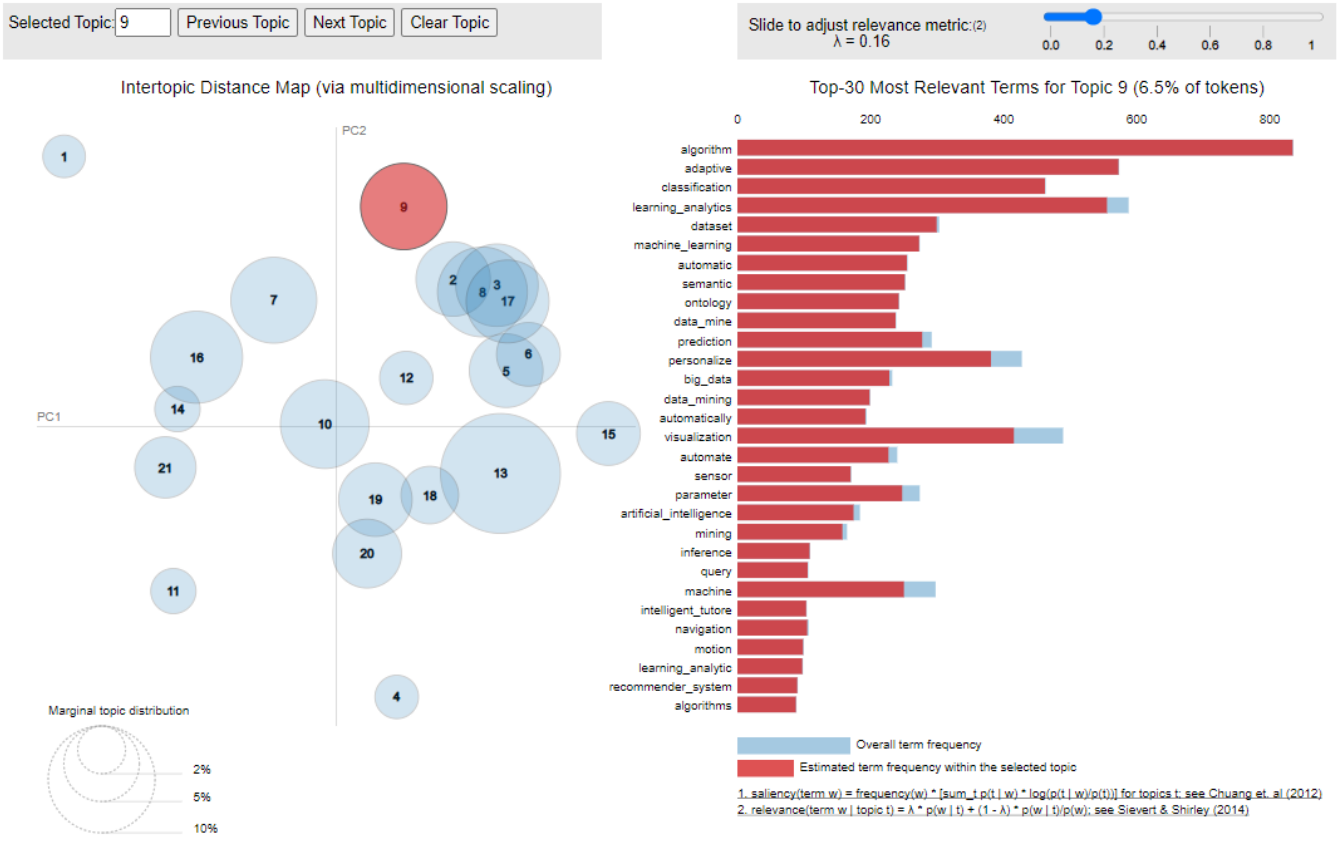


Figura 6-2. *visualización interactiva del paquete pyLDAvis con λ=0.16*

(cada burbuja representa un tópico el cual es descrito por las palabras que se encuentra a la derecha del gráfico)

6.2 Resultados para número de tópicos igual a 18

Cómo se puede observar en la figura 6-3. La distribución espacial de los tópicos genera 5 grandes grupos. A pesar de que muchos tópicos están superpuestos en la gráfica de reducción de dimensionalidad (PCA), los términos que componen cada tópico permiten identificar la semántica particular de cada uno de ellos. Para comprender mejor esto, en la Figura 6-4., se muestran los términos que hacen parte de los tópicos 3, 5, 8, 13, 15, y 17. Los cuales aparecen superpuestos en la Figura 6-3. Se puede evidenciar que, dentro de las 10 principales palabras de cada tópico, existe una diferencia tal que, permite que cada uno de estos se interpretado individualmente, tal como se hará más adelante.

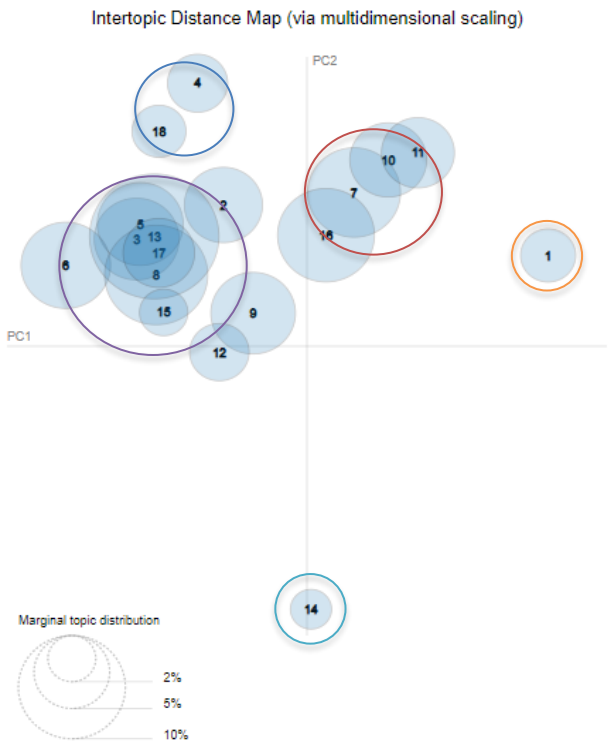


Figura 6-3. *Mapa de distancia entre tópicos*

Figura 6-4. *principales 10 palabras que componen los tópicos 3, 5, 8, 13, 15, y 17*

Otro resultado interesante, es la distribución de la cantidad de documentos recuperados por cada tópico (ver Tabla 8), este valor resulta de gran importancia, principalmente en la contextualización del tema global (Tecnología en la educación) en subtemas de interés. En este caso, para la selección de 18 como valor para el

numero de tópicos, podemos observar cómo los primeros 4 tópicos, agrupan más del 50% de los documentos y como los últimos 7 tópicos agrupan solo el 10%. Estos resultados nos van a permitir al final del ejercicio apoyar a tomar la decisión sobre cuál es el número de tópicos que desde la interpretación humana es más conveniente.

Así mismo, la interpretación de los tópicos, para la cual fue necesario el apoyo de los expertos temáticos (ver Tabla 8), es el resultado de mayor valor, toda vez apunta directamente con el objetivo general de este documento. En este caso 4 tópicos (17, 10, 12, 1) tuvieron que ser revisados más profundamente para poder llegar a una interpretación adecuada.

Tabla 8 *Resumen de información por tópicos para un numero de tópicos igual a 18.*

No. Tópico	Nombre	Cant. Doc.	% Corpus
13	Docente y enseñanza	3195	21,14%
8	Tecnologías virtuales en diseño e ingeniería	1796	11,88%
7	Entrenamiento médico y salud	1423	9,42%
16	Aula invertida, enseñanza a través de juegos y videos	1374	9,09%
9	Sistemas, datos, algoritmos y analítica	1074	7,11%
6	Cursos en línea masivos y abiertos	1060	7,01%
5	Blended learning	920	6,09%
3	Bibliotecas servicios y recursos de información	787	5,21%
2	Modelos de e-learning y sistemas de gestión de aprendizaje (LMS)	713	4,72%
17	* Enseñanza relacionada a redes, multimedia y computación en nube	617	4,08%
11	Motivación y control de grupos académicos y desempeño en clase	542	3,59%
10	* Análisis del comportamiento y del genero	451	2,98%
18	Dispositivos móviles	347	2,30%
4	Tecnologías de información y comunicación (TIC)	291	1,93%
12	* Herramientas de investigación, libros, revistas y bases de datos	247	1,63%
1	* Evaluaciones	127	0,84%
15	Idioma ingles y segundas lenguas	119	0,79%
14	Estilos y espacios de aprendizaje	30	0,20%

* Estos nombres se revisaron disminuyendo el valor λ de la visualización con *pyLDAvis* con el fin de encontrar palabras más específicas pertenecientes a cada tópico.

Finalmente, y con el propósito de profundizar más en el contexto general de los datos obtenidos, se revisó la probabilidad de los 10 principales términos de cada tópico (ver Figura 6-5) y su porcentaje acumulado (ver Figura 6-6). En el primero de los dos análisis, se tiene como objetivo medir la “fuerza” de una palabra dentro de cada tópico, lo cual si las frases tienen suficiente valor semántico nos puede ayudar a identificar que tan dispersos sintácticamente son los documentos; para el segundo análisis, se tiene como objetivo, revisar que tanto “definen” las 10 primeras palabras a un tópico en particular, como medida de referencia entre estos valores y la interpretación final desde el humano.

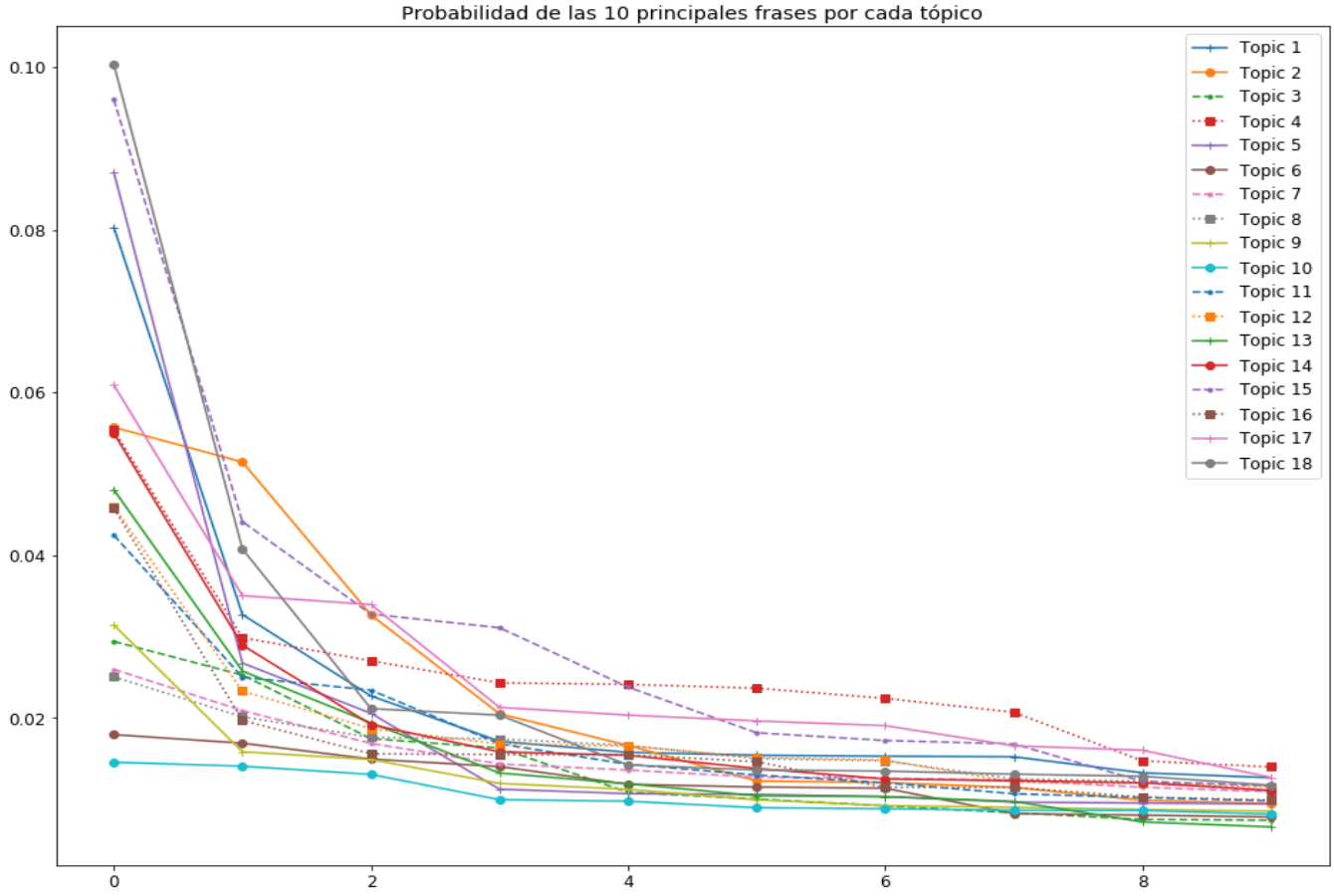


Figura 6-5. *Probabilidad de las 10 palabras que componen cada tópico*

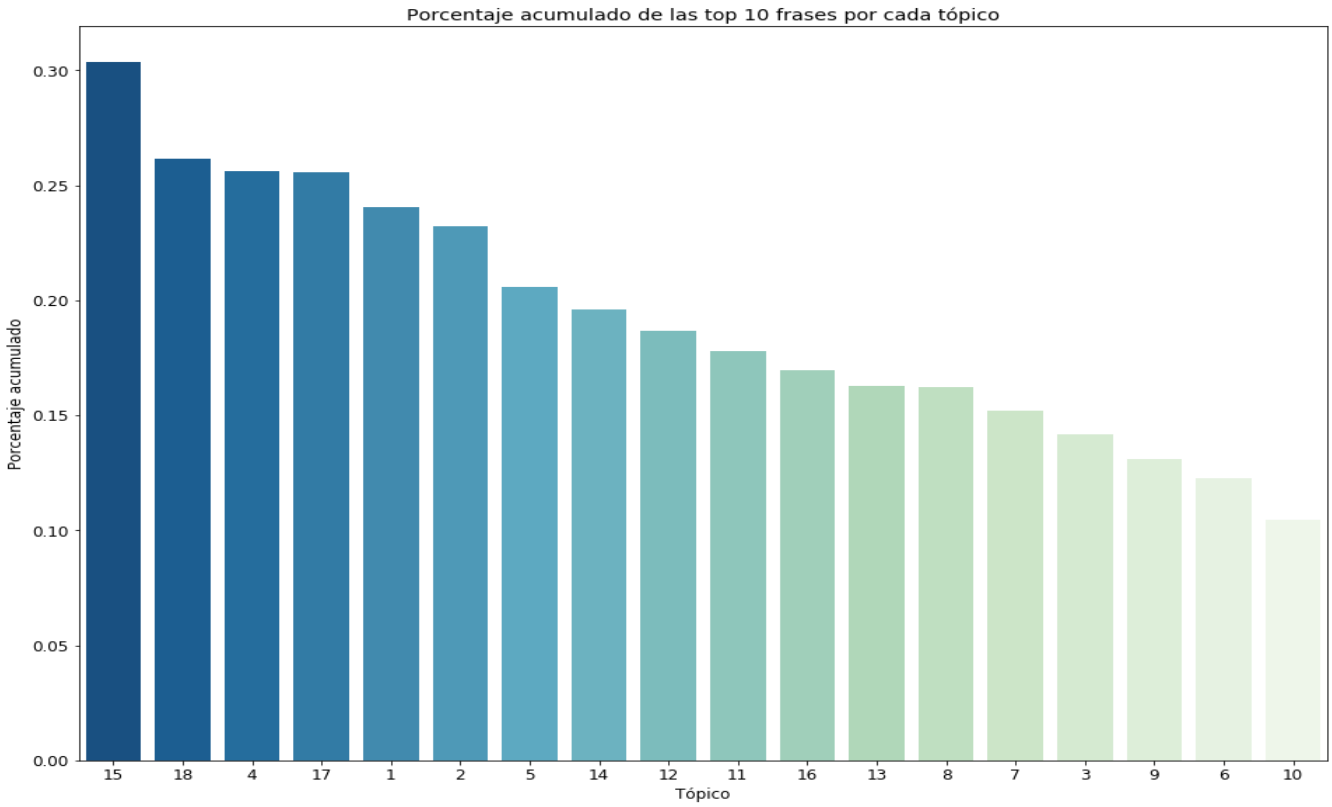


Figura 6-6. *Porcentaje acumulado de las 10 principales palabras que componen cada tópico*

6.3 Resultados para número de tópicos igual a 21

A diferencia del gráfico de mapa de distancia entre tópicos para un numero de tópicos igual a 18, en este caso hay una mejor distancia entre tópicos y una mayor cantidad de sub-agrupaciones tal como se puede observar en la figura 6-7. Teóricamente, esto permite una mejor definición de los tópicos y por lo tanto una mayor cantidad de conocimiento a descubrir debido al aumento de la cantidad de tópicos. Al igual que en el caso anterior se revisó la distribución de la cantidad de documentos recuperados por cada tópico, y en este caso, para la selección de 21 como valor para el numero de tópicos, podemos observar cómo los primeros 5 tópicos, agrupan más del 50% de los documentos y como los últimos 9 tópicos apenas sobrepasan el 10% (ver Tabla 9).

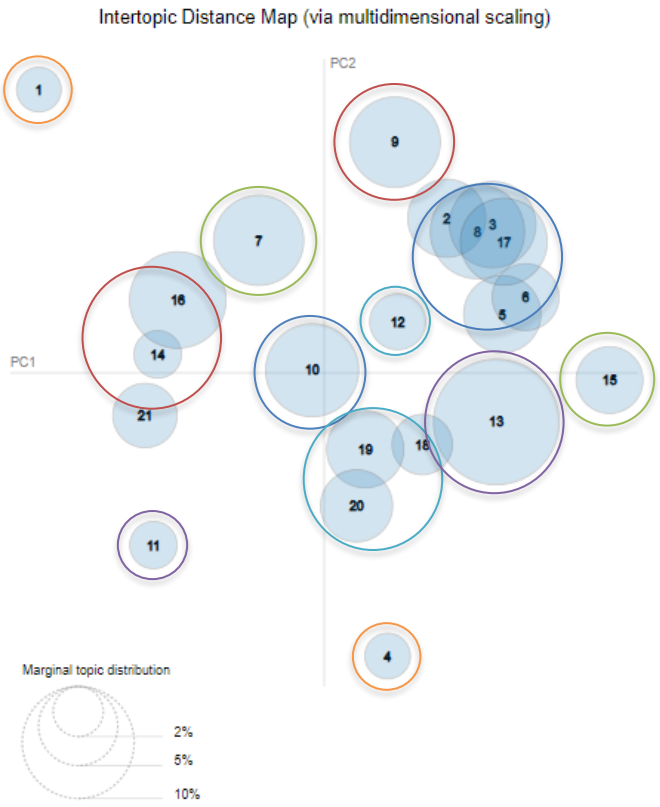


Figura 6-7. *Mapa de distancia entre tópicos*

A continuación, se definieron los nombres (ver Tabla 9) con el acompañamiento de los expertos, se revisaron las palabras que conforman cada tópico y se notó una leve mejoría en cuanto a la interpretación de los tópicos, así como la aparición de grupos de temas que anteriormente no se encontraban definidos y otros que ya estaban definidos, pero que se separaron del tópico principal, con la suficiente cantidad de documentos para generar interés desde la especificidad temática. En este caso 3 tópicos (1, 10,14) tuvieron que ser revisados más profundamente para poder llegar a una interpretación adecuada.

Tabla 9 *Resumen de información por tópicos para un numero de tópicos igual a 21.*

No. Tópico	Nombre	Cant. Doc.	% Corpus
13	Prácticas pedagógicas en espacios y experiencia en el salón de clases	2770	18,3%
8	Tecnologías virtuales en diseño e ingeniería	1361	9,0%
9	Sistemas, datos, algoritmos y analítica	1206	8,0%
7	Entrenamiento médico y salud	1192	7,9%
16	Videos como método educativo	1105	7,3%
10	*Universidad, Competencias y habilidades	1027	6,8%
17	Tecnologías de información y comunicación (TIC)	983	6,5%
19	Aula invertida, enseñanza a través de juegos y videos	870	5,8%
3	Bibliotecas servicios y recursos de información	772	5,1%
2	Modelos de e-learning y sistemas de gestión de aprendizaje (LMS)	691	4,6%
20	Docente y enseñanza	590	3,9%
5	Blended learning	578	3,8%
6	Cursos en línea masivos y abiertos	531	3,5%
18	Dispositivos móviles	432	2,9%
15	Tecnología digital y educación	257	1,7%
21	Comportamiento, factores sociales y psicológicos	225	1,5%
12	Herramientas de investigación, libros, revistas y bases de datos	194	1,3%
11	STEM (Ciencia, Tecnología, Ingeniería y Matemáticas)	94	0,6%
4	Idioma ingles y segundas lenguas	89	0,6%
1	* Evaluaciones	75	0,5%
14	*Retroalimentación	71	0,5%

* Estos nombres se revisaron disminuyendo el valor λ de la visualización con *pyLDAvis* con el fin de encontrar palabras más específicas pertenecientes a cada tópico.

Finalmente, y como para el caso anterior, se revisó la probabilidad de los 10 principales términos por cada tópico (ver Figura 6-8) y el porcentaje acumulado de las 10 principales frases por cada tópico (ver Figura 6-9). Para esta solución, es posible identificar que, dados los nuevos tópicos generados, algunas palabras llegan a tener probabilidades cercanas al 20% dentro de un tópico, identificando 5 tópicos cuya principal palabra representa más del 8%, y cuando se analiza el porcentaje acumulado, la combinatoria de las 10 principales palabras para 13

tópicos sobrepasa el 20% llegando hasta casi el 40% para el que mayor valor acumula.

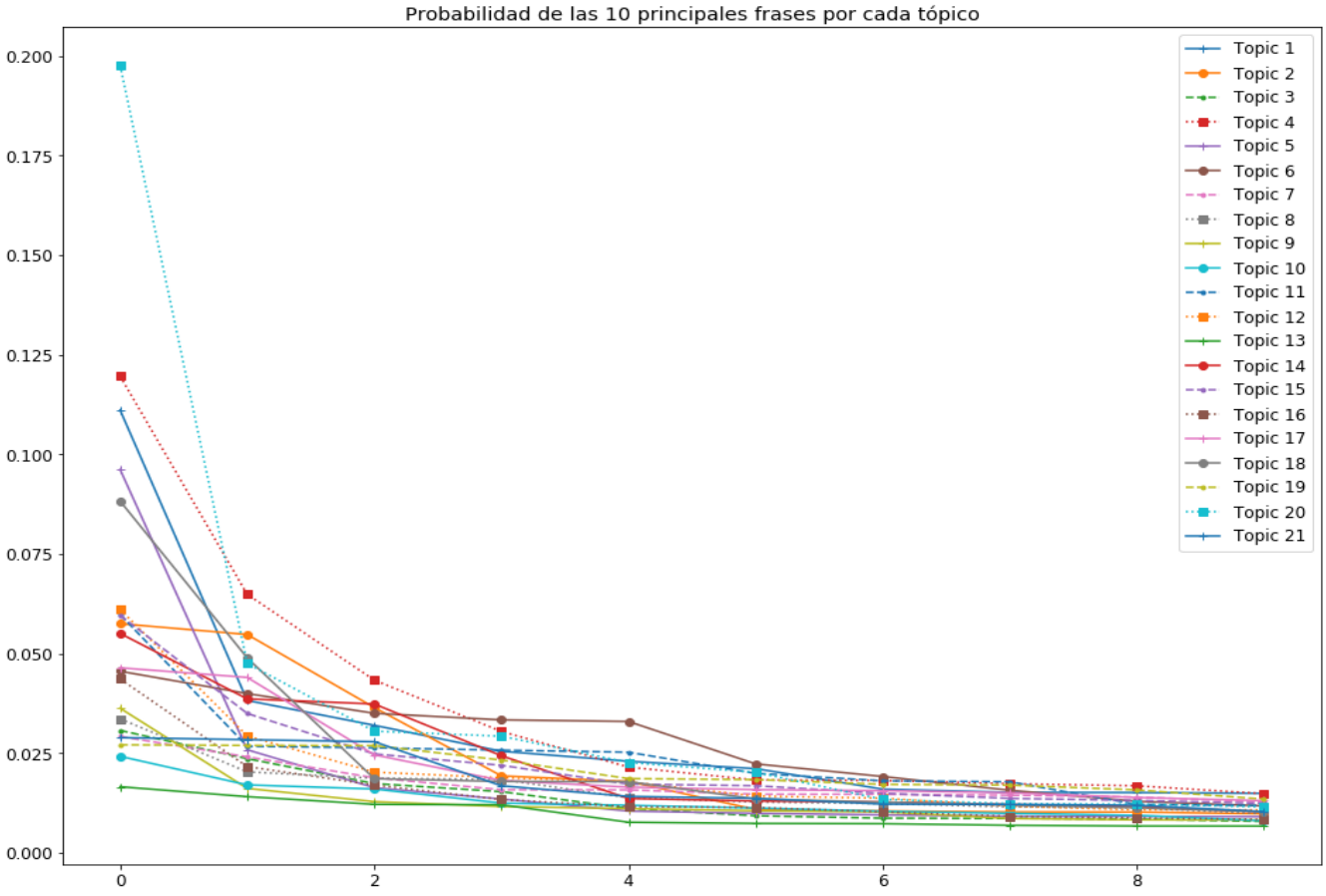


Figura 6-8. *Probabilidad de las 10 palabras que componen cada tópico*

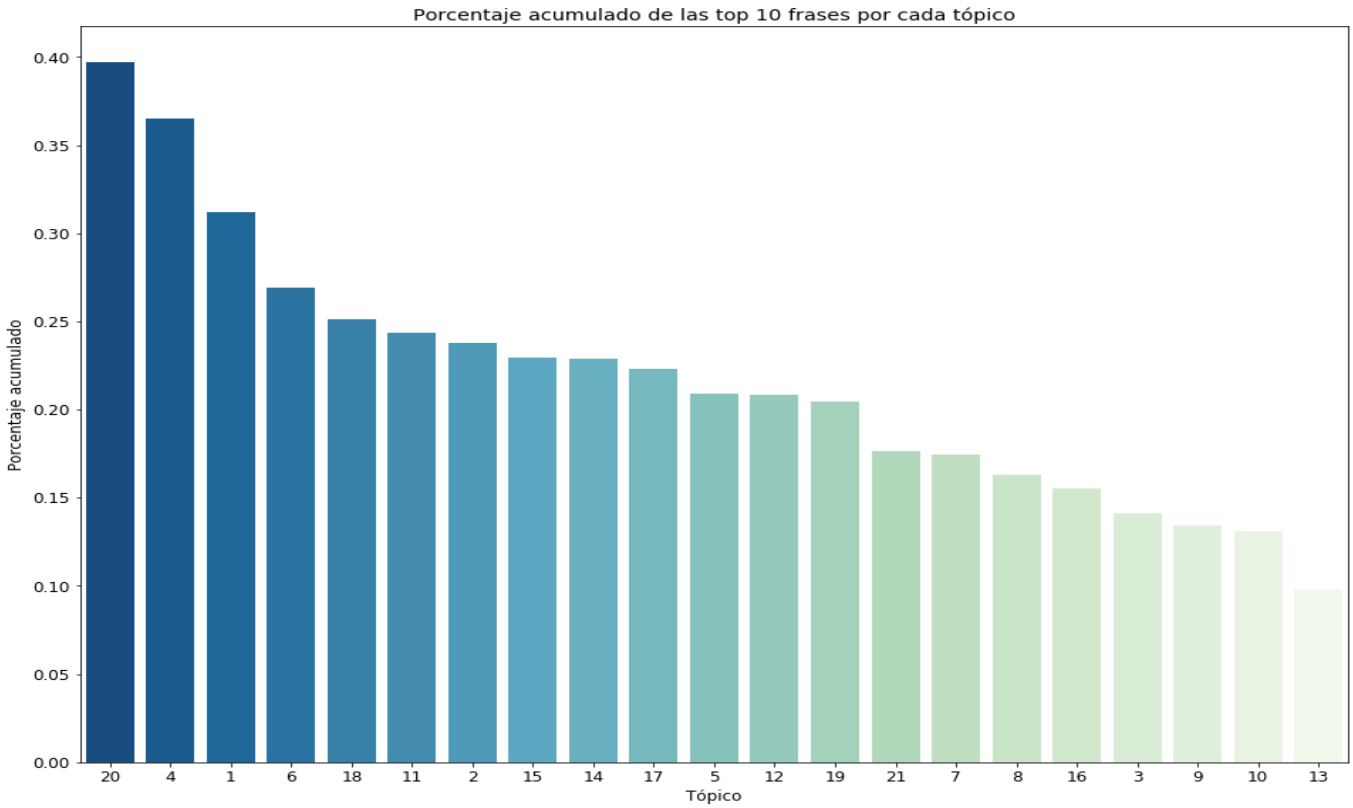


Figura 6-9. *Porcentaje acumulado de las 10 principales palabras que componen cada tópico*

6.4 Resultados para número de tópicos igual a 24

Como último ejercicio, se revisó el resultado de la agrupación de tópicos para un valor de 24. De lo anterior se obtuvo que, el mapa de distancia entre tópicos (ver Figura 6-10.) respecto al mapa para el número de tópicos anterior, genera una mayor cantidad de agrupaciones, tal como se esperaba.

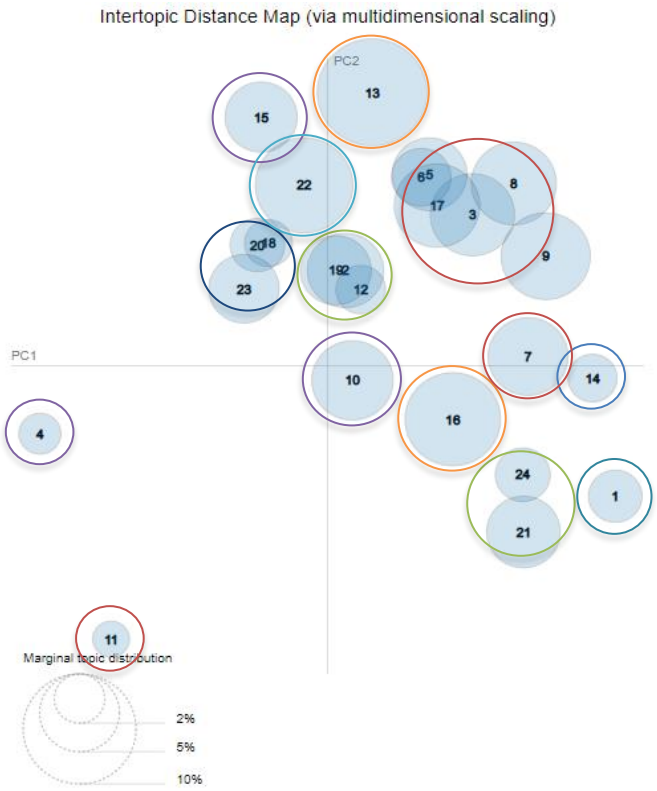


Figura 6-10. *Mapa de distancia entre tópicos*

Respecto a la cantidad de documentos que se encuentran agrupados, se halló lo siguiente: primero, que los primeros 6 tópicos agrupan más del 50% de los documentos y segundo, que al igual que en el caso anterior los últimos 9 tópicos (ver Tabla 10) apenas sobrepasan el 10%. Para el primer caso, los resultados coinciden con la creación de un nuevo tópico, consecuencia de la separación del tópico más grande en 2 nuevos tópicos; y para el segundo se mantiene la tendencia de los resultados anteriores.

Para mantener la consistencia del ejercicio, se revisó la probabilidad de los 10 principales términos por cada tópico (ver Figura 6-11.) y el porcentaje acumulado de estos (ver Figura 6-12.). Para esta solución, es posible identificar que, algunas palabras llegan a tener probabilidades de un poco más del 14% y que 5 tópicos siguen teniendo en su palabra principal porcentajes de más de 8%. Es importante

mencionar que, si bien para el análisis de cada tópico individualmente este dato es importante, este valor por sí solo, no indica una desmejora respecto al caso anterior en donde se alcanzaban porcentajes de casi el 20%; Por su parte, cuando se analiza el porcentaje acumulado, la combinatoria de las 10 principales palabras para 16 tópicos sobrepasa el 20% llegando hasta un poco más del 35% para el que mayor valor acumula.

Finalmente, se procedió a revisar la facilidad de interpretación de los tópicos generados utilizando el acompañamiento de los expertos. En este caso, 8 tópicos (13, 16, 17, 23, 20, 1, 4, 14) tuvieron que ser revisados más profundamente para poder llegar a una interpretación adecuada.

Tabla 10 *Resumen de información por tópicos para un numero de tópicos igual a 24.*

No. Tópico	Nombre	Cant. Doc.	% Corpus
13	*Espacios de aprendizaje	1836	12,15%
22	Docente y enseñanza de idiomas	1627	10,77%
16	*Espacios(salón) de clases	1208	7,99%
9	Sistemas, datos, algoritmos y analítica	1173	7,76%
8	Tecnologías virtuales en diseño e ingeniería	1143	7,56%
17	* Enseñanza relacionada a redes, multimedia y computación en nube	991	6,56%
7	Entrenamiento de enfermería y salud	734	4,86%
19	Aula invertida, enseñanza a través de juegos y videos	720	4,76%
3	Bibliotecas, servicios y recursos de información	718	4,75%
2	Modelos de e-learning y sistemas de gestión de aprendizaje (LMS)	693	4,59%
21	intervención médico-paciente y salud	674	4,46%
10	Tecnologías de información y comunicación (TIC)	660	4,37%
5	Blended learning	548	3,63%
6	Cursos en línea masivos y abiertos (Moocs)	445	2,94%
23	*Educación previa a la enseñanza (Pre-service teacher)	412	2,73%
15	Tecnología digital y educación	411	2,72%
18	Dispositivos móviles	262	1,73%
24	Videos como método educativo	225	1,49%
20	*Enseñanza de matemáticas, deporte, música, ciencias sociales y naturales	206	1,36%

12	Competencias informacionales	125	0,83%
1	* Evaluaciones	120	0,79%
4	*Adaptación a la tecnología	104	0,69%
14	*Materiales y contenidos	66	0,44%
11	STEM (Ciencia, Tecnología, Ingeniería y Matemáticas)	12	0,08%

* Estos nombres se revisaron disminuyendo el valor λ de la visualización con *pyLDAvis* con el fin de encontrar palabras más específicas pertenecientes a cada tópico.

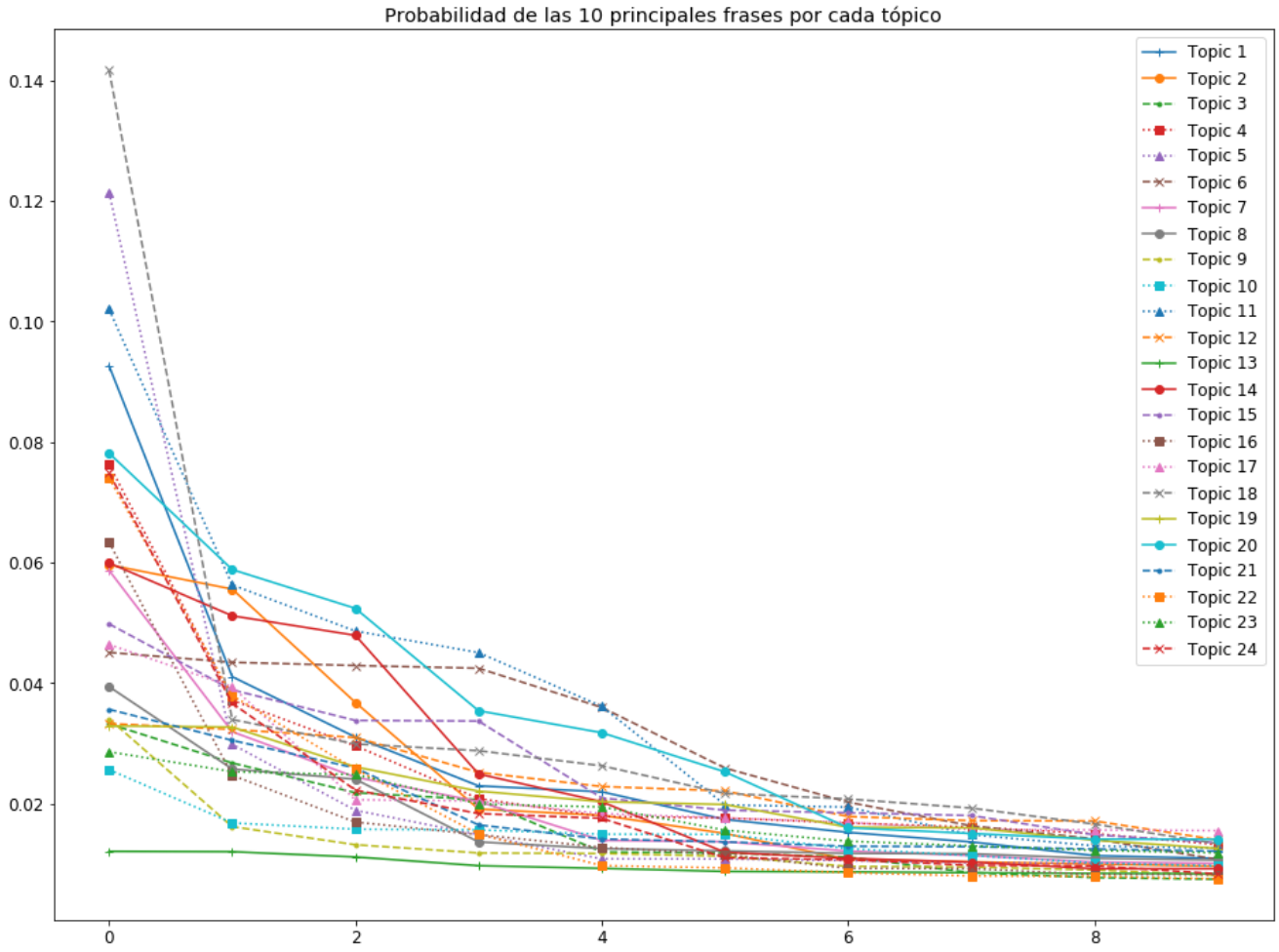


Figura 6-11. *Probabilidad de las 10 palabras que componen cada tópico*

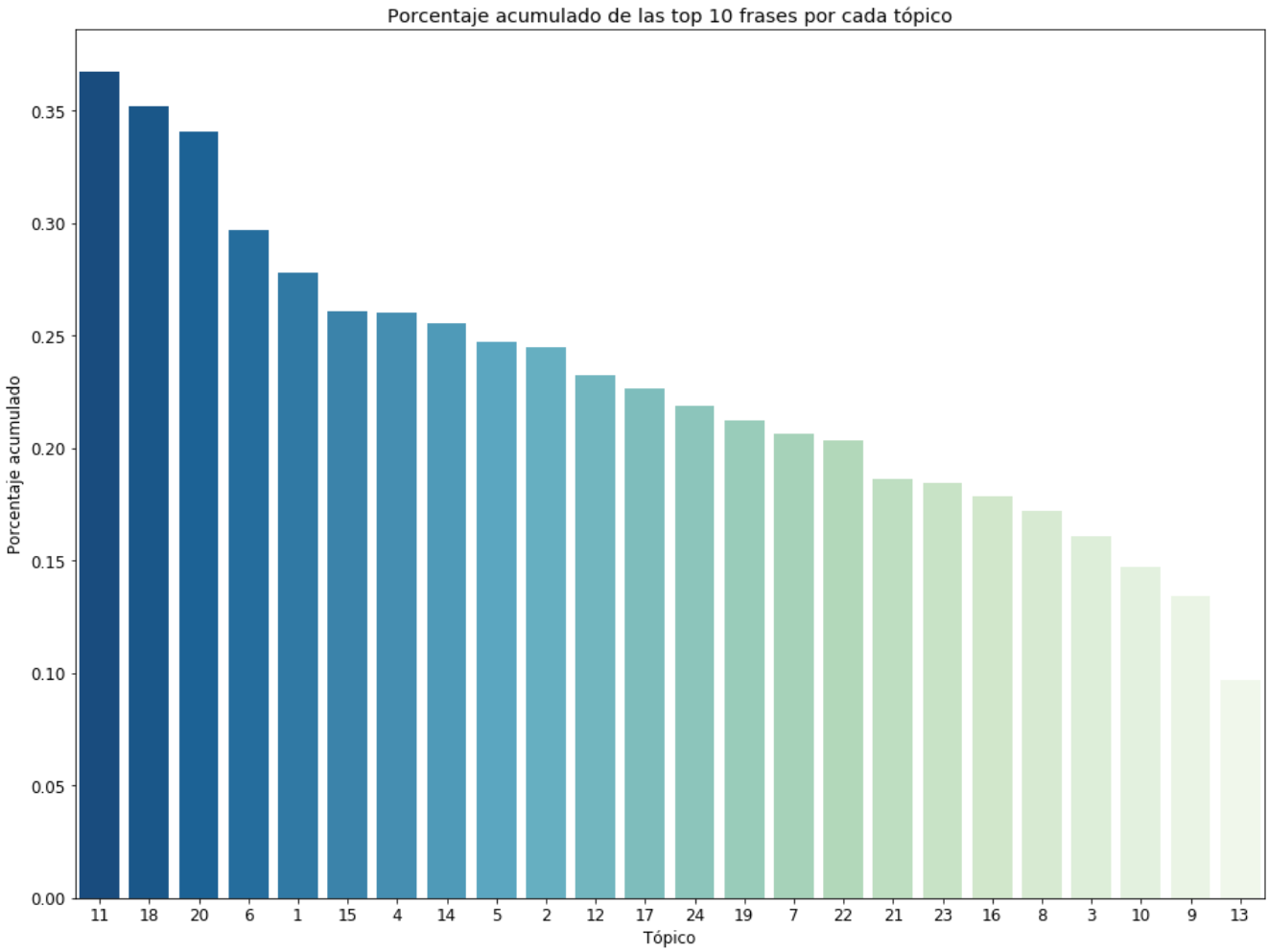


Figura 6-12. *Porcentaje acumulado de las 10 principales palabras que componen cada tópico*

6.5 Selección del modelo con mejor interpretabilidad desde la perspectiva humana

Después de realizar el ejercicio de generar los tópicos e interpretarlos, se procedió a realizar una evaluación en la que se valoraron 2 variables por parte de 2 expertos. La primera fue la facilidad para interpretar cada tópico y la segunda la capacidad de descubrimiento de conocimiento. Para esto, se solicitó que se calificara de 1 a 10 las anteriores variables siendo 1 el valor de mayor dificultad y 10 el valor de mayor facilidad. Debido a que, para el ejercicio, la interpretación de los tópicos es de mayor relevancia, se le asignó un peso del 60% a esta variable y un 40% a el descubrimiento de conocimiento los resultados de pueden ver en la tabla 11.

Tabla 11 *Matriz para la selección del mejor modelo.*

Variable	Peso	Experto	No. Tópicos 18	No. Tópicos 21	No. Tópicos 24
Facilidad interpretación	30%	1	7	8	7
Descubrimiento de conocimiento	20%	1	7	8	8
Facilidad interpretación	30%	2	7	8	6
Descubrimiento de conocimiento	20%	2	7	7	8
Total Calificación	**100%**		**7**	**7.8**	**7**

Finalmente, con base en los resultados de la matriz, se seleccionó como modelo a trabajar el modelo con valor 21 para el parámetro número de tópicos.

6.6 Análisis descriptivos con base en la información disponible

Tal como se mencionó al inicio de este capítulo, gracias a que los datos lo permiten, se va a relacionar la base de datos original con la base de datos resultado del análisis de tópicos. Esto con el fin de evaluar, cuál ha sido el comportamiento de los temas en el tiempo y adicionalmente, desagregar esta información por *Journals* académicos. Es importante que se tenga en cuenta que estos datos corresponden a los resultados generados después de realizar la búsqueda “Tecnología en la educación”, pues el hecho que a través de la búsqueda se descubran resultados como por ejemplo, MOOC, LSM, STEM, no implica que lo que se descubre, sea el universo de cada uno de los tópicos. Por lo tanto, si el interés es un análisis global de cada uno de los temas, se debería recurrir de nuevo a la búsqueda de información desde el agregador bibliográfico.

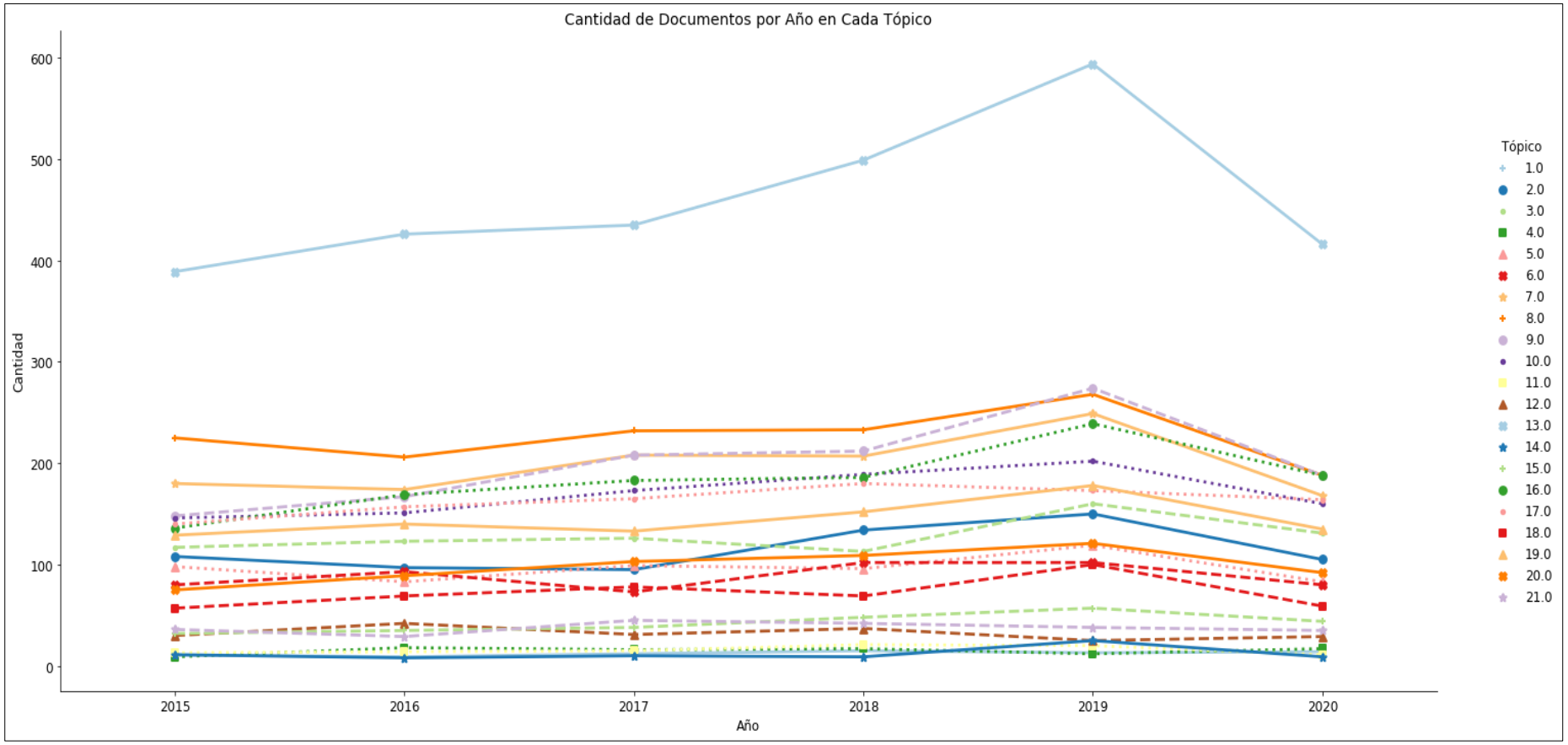


Figura 6-133. *Comportamiento de publicaciones de los Tópicos durante el periodo enero 2015 - agosto 2020*

A continuación, se analizó el comportamiento de la cantidad de documentos publicados durante el periodo de evaluación (ver Figura 6-13.) y se identificaron algunos datos importares tales como:

- El tópico 13 que corresponde a “Prácticas pedagógicas en espacios y experiencia en el salón de clases” (Ver tabla 9) ha tenido un crecimiento sostenido. En este tópico los espacios hacen referencia no solo al aula de clase, sino a todo espacio físico en el que el estudiante puede interactuar incluye museos, bibliotecas, espacios abiertos entre otros.
- El tópico 9 (Sistemas, datos, algoritmos y analítica) al igual que el 16 (Videos como método educativo) y otros como el tópico 2 (Modelos de e-learning y sistemas de gestión de aprendizaje (LMS)) también presentan crecimientos sostenido importantes lo que se puede interpretar como un interés en general de buscar métodos alternativos a método tradicional de educación y el atractivo que en los últimos tiempos han tenidos los temas de los sistemas, datos, algoritmos y analítica y su contexto dentro de la educación.
- El tópico 19 (Aula invertida, enseñanza a través de juegos y videos) y el 8 (Tecnologías virtuales en diseño e ingeniería) también es consistente con las apreciaciones anteriormente mencionadas. En este sentido, se nota el aumento en el interés en estos temas de investigación relacionados con la educación. Temas como la realidad virtual los videojuegos hacen parte de estas tendencias.

Para todos los datos, es importante mencionar que el año 2020 es atípico, toda vez que la cantidad de información generada solo esta hasta el mes de agosto y que, debido a la situación de pandemia, es posible que muchas universidades y centros de investigación hayan disminuido significativamente su producción académica.

Teniendo en cuenta lo anteriormente mencionado, el análisis descriptivo que se logra generar a través del anterior análisis esta información está orientado a ayudar a obtener una vista general de la situación actual de la investigación en particular

de este proyecto, por lo que se insiste, en que el comportamiento de los tópicos sobre los que la cantidad de documentos no permite visualizar cambios significativos, no deberían de ser tenidos en cuenta para interpretación alguna más allá de la existencia del tópico

Otro de los análisis que vale la pena llevar a cabo dada la información existente, es que revistas (*Journals*) *están* abordando cada uno de los tópicos y en general que revistas abarcan el contexto global (ver Figura 6-14.). Esto es importante al momento de realizar los análisis cienciométricos respectivos previos a iniciar la fase exploratoria y sobre todo en el momento de publicar cuando se está buscando generar la mayor visibilidad e impacto del producto de investigación, previo a la selección de una revista en la que se desee publicar.

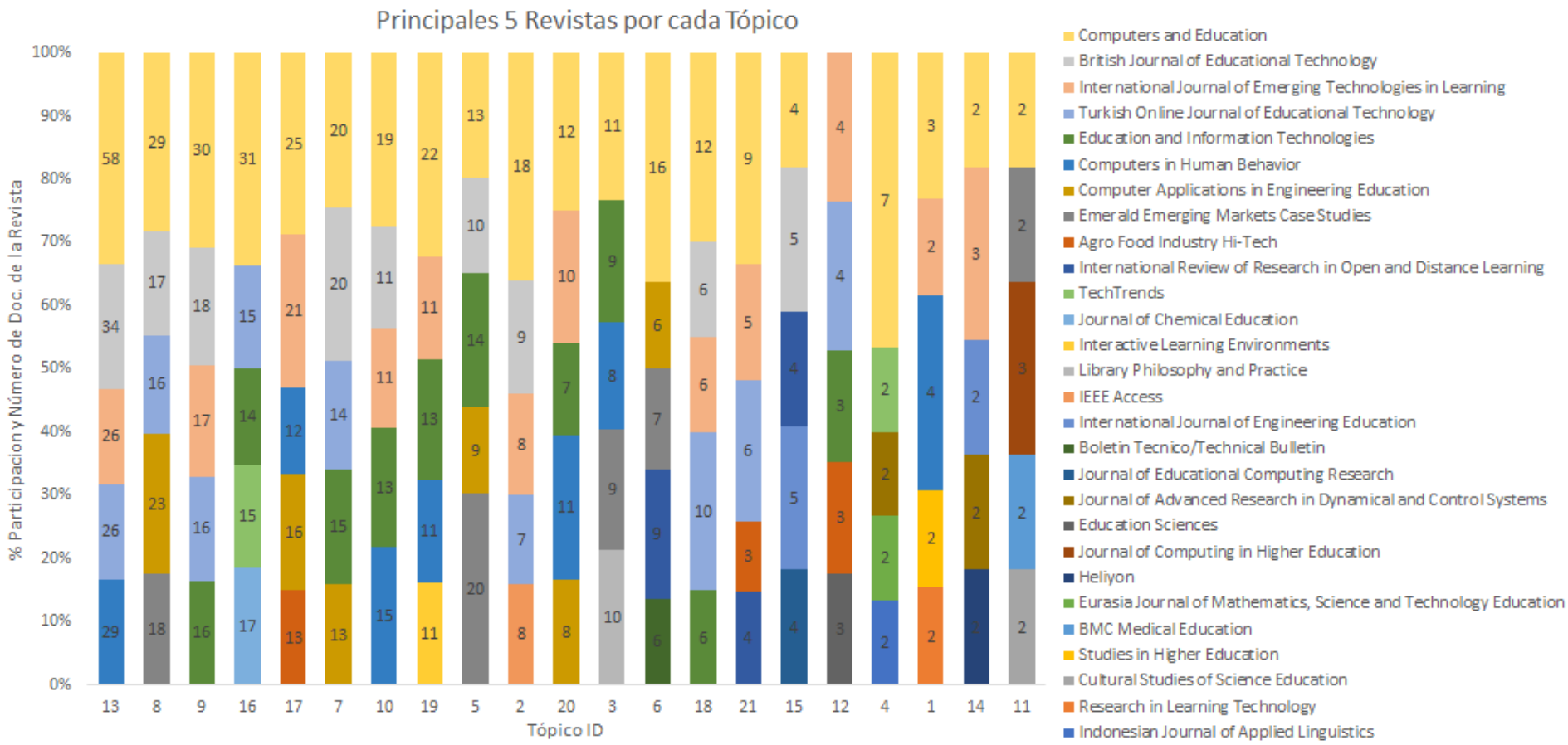


Figura 6-14. *Principales 5 Revistas por cada Tópico*

6.7 Similitud de Documentos

Uno de los resultados más importantes de este trabajo, es la asignación a un tópico de un nuevo documento. Esto es de gran valor en el momento de que se tenga un resumen de un documento, unas palabras clave o una oración y se quiera identificar a qué tópicos pertenece y en qué proporción. Con esta información, se puede optimizar la búsqueda de información al dirigir los esfuerzos de recopilación y análisis a solo los documentos afines, además de aprovechar la información del análisis descriptivo de estos tópicos para los fines que determine el investigador.

Para identificar la pertenencia de un documento a un tópico, utilizamos el método *get_document_topics* de la librería *gensim*. el resultado, es una lista con la probabilidad de que el documento pertenezca a un tópico determinado. Con el fin de simular a manera de ejemplo la situación, se le paso al modelo el siguiente texto:

“In this document we will talk about the importance of academic spaces within the process of teaching the arts. Places such as museums, libraries will be used as interactive learning spaces and will be compared with non-face-to-face media but guided by technology such as virtual reality and interactive videos, to finally relate it to the effectiveness of the classroom.”

Claramente, se diseñó un texto cuya intención era ser relacionado por el algoritmo con el título del tópico 13, lo cual se logró como se puede ver en los resultados de la figura 6-15. En este caso, el valor 12, corresponde al tópico 13, esto dada la particularidad de *Python* de iniciar la numeración en 0. Con este resultado, se da por concluida la fase de modelado.

```
bow = lda_model21.id2word.doc2bow(data_lemmatized[0])#crear bolsa de palabras


#obtener la clasificación del tópico


doc_topics = lda_model21.get_document_topics(bow, per_word_topics=True)


print(doc_topics)

[(2, 0.0787441), (7, 0.10328662), (12, 0.30809867), (14, 0.15842354), (15, 0.15701805),
(16, 0.14535081)]
```

Figura 6-15. *Resultado de clasificación de un documento nuevo*

7. VALIDACIÓN

Uno de los objetivos del presente trabajo es evaluar la calidad y pertinencia del modelo desarrollado en la determinación de temas de investigación relevantes. Para verificar lo anterior, se procedió a realizar un muestreo estratificado (ver Anexo A) cuyo número óptimo de muestra seleccionada fue de 385.

Para cada grupo, se calculó el tamaño de muestra (ver Tabla 12), la cual fue aleatoriamente seleccionada y posteriormente evaluada por el experto temático. El ejercicio de validación consistió en leer el resumen del articulo y determinar si este, desde la interpretación humana, coincide o no con el nombre del tópico asignado.

Tabla 12 *Selección del Tamaño de muestra para cada grupo temático*

Grupo	Ni	ni
13	2770	69
8	1361	34
9	1206	30
7	1192	30
16	1105	28

10	1027	26
17	983	25
19	870	22
3	772	20
2	691	18
20	590	15
5	578	15
6	531	14
18	432	11
15	257	7
21	225	6
12	194	5
11	94	3
4	89	3
1	75	2
14	71	2
Total	**15113**	**385**

Los resultados obtenidos muestran, que de 385 documentos que se revisaron, 267 se relacionan con el nombre del tópico propuesto (ver Tabla 13) lo que corresponde a un valor de 69,35% para la correcta clasificación del modelo de manera global.

Tabla 13 *Resultados de la evaluación del experto, referente a la correcta clasificación de los documentos en cada muestra*

Grupo	ni	Clasificados correctamente
13	69	45
8	34	23
9	30	20
7	30	21
16	28	19
10	26	18
17	25	18
19	22	15
3	20	13
2	18	13
20	15	11

5	15	10
6	14	10
18	11	9
15	7	5
21	6	5
12	5	4
11	3	2
4	3	2
1	2	2
14	2	2
Total	**385**	**267**

Los resultados obtenidos, sugieren que en general el algoritmo LDA, resulta ser de bastante utilidad durante la determinación de temas de investigación relevantes. Esto, a partir de la capacidad que tiene la herramienta para lograr generar un contexto global sobre los tópicos contenidos en un conjunto de documentos. en el este caso particular de este trabajo y dentro del área de interés "*Tecnología en la educación"* los expertos temáticos, resaltan la utilidad de la herramienta como herramienta de apoyo durante la fase de exploratoria previa a el desarrollo de un trabajo de investigación.

8. CONCLUSIONES Y TRABAJOS FUTUROS

Durante la planeación del proyecto de grado de la Maestría en Ciencia de Datos, se vio la oportunidad de acercar la Ciencia de Datos a la solución de situaciones reales que ocurren en contextos académicos relacionados a la investigación. En este sentido este trabajo resulta muy enriquecedor desde el aprendizaje obtenido, no solo a partir de desarrollo del modelo, sino también desde el perfeccionamiento de competencias informacionales que se crearon durante el desarrollo del presente trabajo. Igualmente, con el acercamiento al tema sobre el que se desarrolló el modelo, se logró entender de manera global un contexto temático desconocido por el autor hasta el momento y que, dado el momento actual de la historia, cobra gran validez en cuanto a la necesidad de adaptar nuevas metodologías y tecnologías en el desarrollo de la educación.

Este trabajo muestra que es posible utilizar herramientas de *machine learning* para determinar temas de investigación, con el fin de apoyar el proceso de investigación durante la fase inicial de exploración, no solo desde el área temática trabajada, sino desde cualquier tema de interés. Lo anterior teniendo en cuenta que los resultados de este trabajo se pueden replicar fácilmente a otros campos del conocimiento e incluso a otro tipo de contextos completamente diferentes.

Teniendo en cuenta los objetivos propuestos, a continuación, se concluye sobre cada uno de ellos:

El proceso de recolección de información desde bases de datos académicas es un proceso que, si bien parece sencillo, presenta grandes desafíos. Lo anterior debido a la limitada información que estas bases de datos brindan a los usuarios en una sola consulta y a lo costoso que el acceso a estas herramientas resulta para las instituciones académicas. En general el acceso a los metadatos, no resulta ser un

proceso sencillo para el usuario, requiere de conocimiento especializado en estas herramientas y de procesos de transformación de datos para lograr realizar análisis de estos metadatos.

Aplicar técnicas de procesamiento de lenguaje natural, requiere de un gran esfuerzo en el preprocesamiento de la información. Durante este proceso se pueden utilizar múltiples técnicas de limpieza, pero también es necesario contar con el apoyo de los expertos temáticos quienes al final cumplen un papel muy importante en la definición del conjunto de palabras que van a ser ingresadas al diccionario.

El balance entre la calidad del modelo y la disponibilidad de tiempo es un factor bastante importante a tener en cuenta en este tipo de trabajos. La identificación de los hiper-parámetros con lo que se obtienen mejores indicadores fue una de las partes del proyecto que más “tiempos muertos” generó. Finalmente, no se logró concluir que haber optimizado los hiper-parametros haya influido en una mejor agrupación e interpretación de los Tópicos.

Tal como se encontró en la literatura, las medidas de bondad del modelo no reemplazan la interpretación humana. El modelo LDA es un modelo probabilístico que, si bien tiene buenos resultados, no interpreta semánticamente los documentos por lo que en ocasiones crea tópicos los cuales pueden tener sentido desde el valor numérico, pero carecen de sentido desde la interpretación humana.

El resultado de la validación de este trabajo fue destacado por los expertos temáticos, quienes después de revisar los documentos y las agrupaciones correspondientes, resaltan la utilidad de la herramienta como herramienta de apoyo durante la fase de exploratoria previa a el desarrollo de un trabajo de investigación.

Finalmente se proponen trabajos que se fueron divisando durante el desarrollo de este proyecto, pero que no hacen parte del alcance:

La implementación del modelo LDA a partir de computación distribuida usando GPU´s puede ayudar a mejorar los tiempos de entrenamiento del modelo y en la búsqueda de mejores hiper-parámetros.

La comparación entre los resultados obtenidos a través de librerías existentes que abordan el modelo LDA. Puede ser una posibilidad adicional en el momento de escoger el modelo que mejor se ajuste a la interpretación humana.

El desarrollo de una aplicación que permita utilizar esta herramienta como apoyo a procesos de vigilancia estratégica puede tener grandes ventajas en cuanto al descubrimiento de información al momento de la exploración de documentos de texto que se quieran evaluar.

BIBLIOGRAFÍA

Advanced Analytical Theory and Methods. (2015). *Data Science & Big Data Analytics*, 255–293. https://doi.org/10.1002/9781119183686.ch9

Barker, K., & Cornacchia, N. (2000). *Using Noun Phrase Heads to Extract Document Keyphrases* (pp. 40–52). https://doi.org/10.1007/3-540-45486-1_4

Blei, D. M., Ng, A. Y., & Edu, J. B. (2003). Latent Dirichlet Allocation Michael I. Jordan. In *Journal of Machine Learning Research* (Vol. 3).

Castillo, C. (2019). *Implementación De Técnicas Para Minería De Texto Usando Modelos De Tópicos*. http://bibdigital.epn.edu.ec/handle/15000/19998

Cerratto Pargman, T., & Jahnke, I. (Eds.). (2019). *Emergent Practices and Material Conditions in Learning and Teaching with Technologies*. Springer International Publishing. https://doi.org/10.1007/978-3-030-10764-2

Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. *Advances in Neural Information Processing Systems 22 - Proceedings of the 2009 Conference*, 288–296.

Dinero. (2018). *Ranking de las empresas más innovadoras de Colombia en 2018*. Dinero. https://www.dinero.com/edicion-impresa/caratula/articulo/ranking-de-las-empresas-mas-innovadoras-de-colombia/246812

Edquist, H., & Henrekson, M. (2017). Do R&D and ICT affect total factor productivity growth differently? *Telecommunications Policy*, *41*(2), 106–119. https://doi.org/10.1016/j.telpol.2016.11.010

Fernández, L. A. U. (2019). *Reducir el número de palabras de un texto: lematización y radicalización (stemming) con Python*. https://medium.com/qu4nt/reducir-el-número-de-palabras-de-un-texto-lematización-y-radicalización-stemming-con-python-965bfd0c69fa

Ganegedara, T. (2018). *Intuitive Guide to Latent Dirichlet Allocation - Towards Data Science*. https://towardsdatascience.com/light-on-math-machine-learning-intuitive-guide-to-latent-dirichlet-allocation-437c81220158

Greenville, A. C., Dickman, C. R., & Wardle, G. M. (2017). 75 years of dryland science: Trends and gaps in arid ecology literature. *PLOS ONE*, *12*(4), e0175014. https://doi.org/10.1371/journal.pone.0175014

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(SUPPL. 1), 5228–5235. https://doi.org/10.1073/pnas.0307752101

Hernández, E., Tomás, A. ;, & Navarro, D. ; (2015). *Procesamiento del Lenguaje Natural Sociedad Española para el Procesamiento del Lenguaje Natural*. http://www.redalyc.org/articulo.oa?id=515751524010

Hoffman, M. D., Blei, D. M., & Bach, F. (2010). Online learning for Latent Dirichlet Allocation. *Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010, NIPS 2010*.

Hui, J. (2019). *Machine Learning — Latent Dirichlet Allocation LDA*. Medium.Com. https://medium.com/@jonathan_hui/machine-learning-latent-dirichlet-allocation-lda-1d9d148f13a4

Iinuma, M. (2016a). *Learning and Teaching with Technology in the Knowledge Society*. Springer Singapore. https://doi.org/10.1007/978-981-10-0144-4

Iinuma, M. (2016b). *New Literacy, Collaboration, and Technology* (pp. 29–42). https://doi.org/10.1007/978-981-10-0144-4_3

Korshunova, I., Xiong, H., Fedoryszak, M., & Theis, L. (2019). *Discriminative Topic Modeling with Logistic LDA*. https://arxiv.org/abs/1909.01436

Li, X., Xie, Q., Daim, T., & Huang, L. (2019). Forecasting technology trends using text mining of the gaps between science and technology: The case of perovskite solar cell technology. *Technological Forecasting and Social Change*, *146*, 432–449. https://doi.org/10.1016/j.techfore.2019.01.012

Lin, J. (2016). *On The Dirichlet Distribution*.

Luo, L. xia. (2019). Network text sentiment analysis method combining LDA text representation and GRU-CNN. *Personal and Ubiquitous Computing*, *23*(3–4), 405–412. https://doi.org/10.1007/s00779-018-1183-9

Marshall, S. J. (2018). *Shaping the University of the Future*. Springer Singapore. https://doi.org/10.1007/978-981-10-7620-6

Nabli, H., Ben Djemaa, R., & Ben Amor, I. A. (2018). Efficient cloud service discovery approach based on LDA topic modeling. *Journal of Systems and Software*, *146*, 233–248. https://doi.org/10.1016/j.jss.2018.09.069

Ng, W. (2015). *New Digital Technology in Education*. Springer International Publishing. https://doi.org/10.1007/978-3-319-05822-1

Sam Tazzyman, DaSH, M. (n.d.). *NLP-guidance*. https://moj-analytical-services.github.io/NLP-guidance/LDA.html

Sampson, D., Ifenthaler, D., Spector, J. M., & Isaías, P. (Eds.). (2018). *Digital Technologies: Sustainable Innovations for Improving Teaching and Learning*. Springer International Publishing. https://doi.org/10.1007/978-3-319-73417-0

Sarkar, D. (2019). Text Summarization and Topic Models. In *Text Analytics with Python* (2nd ed., pp. 343–451). Apress. https://doi.org/10.1007/978-1-4842-4354-1_6

Srinivasa-Desikan, B. (2018). Topic Models. In *Natural Language Processing and Computational Linguistics: a practical guide to text analysis with Python, Gensim, spaCy and Keras* (pp. 128–143). Packt Publishing Ltd.

Wen, S., Li, Z., & Li, J. (2014). *Enhance Social Context Understanding with Semantic Chunks* (pp. 251–262). https://doi.org/10.1007/978-3-662-45924-9_23

Westgate, M. J., Barton, P. S., Pierson, J. C., & Lindenmayer, D. B. (2015). Text analysis tools for identification of emerging topics and research gaps in conservation science. *Conservation Biology*, *29*(6), 1606–1614. https://doi.org/10.1111/cobi.12605

Witten, I. H., Paynter, G. W., Frank, E., Gutwin, C., & Nevill-Manning, C. G. (1999). KEA. *Proceedings of the Fourth ACM Conference on Digital Libraries - DL '99*, 254–255. https://doi.org/10.1145/313238.313437

Yau, C. K., Porter, A., Newman, N., & Suominen, A. (2014). Clustering scientific documents with topic modeling. *Scientometrics*, *100*(3), 767–786. https://doi.org/10.1007/s11192-014-1321-8

ANEXOS

Anexo A *Formulas para el cálculo de la muestra optima en un muestreo estratificado*

Error de estimación

$$E = \frac{d^2}{{z_{1-\propto/2}}^2}$$

Tamaño total de la muestra

$$n = \frac{\sum_{i=1}^{l} N_i P_i Q_i}{NE + \frac{1}{N}\sum_{i=1}^{l} N_i P_i Q_i}$$

Tamaño de cada estrato

$$n_i = n\left[\frac{N_i}{\sum_{i=1}^{l} N_i}\right] = n\left[\frac{N_i}{N}\right] = n[W_i]$$