

Universidad de La Habana  
Facultad de Matemática y Computación



# **Análisis de la Evolución de Tópicos en Publicaciones Médicas Usando Modelos de Tópicos Dinámicos**

Autor:

**Naomi Lahera Champagne**

Tutor:

**Dr. Luciano García Garrido**

Trabajo de Diploma  
presentado en opción al título de  
Licenciado en Ciencia de la Computación

3 de febrero de 2025

[github.com/naomi-lahera/cfdtm-detm](https://github.com/naomi-lahera/cfdtm-detm)


# Agradecimientos

Quiero expresar mi más profunda gratitud a mi familia, cuyo apoyo incondicional y amor han sido fundamentales en cada momento, a mis amigos, por su compañía y aliento, y a mi tutor, por su guía y paciencia. A todos ustedes, muchas gracias.

# Opinión del tutor

Opiniones de los tutores

# Resumen



Los modelos de tópicos son herramientas esenciales para extraer estructuras temáticas latentes en grandes volúmenes de texto. En este estudio, se evaluaron los modelos de tópicos dinámicos *Dynamic Embedding Topic Model (DETM)* y *Chain-Free Dynamic Topic Model (CFDTM)* sobre un conjunto de datos compuesto por artículos de la Revista de Ciencias Médicas de La Habana, Cuba. El objetivo fue identificar cuál modelo se adapta mejor a los datos, utilizando métricas de coherencia de tópicos, diversidad de tópicos, clustering y clasificación. Tras una serie de experimentos, se determinaron las configuraciones óptimas de hiperparámetros para ambos modelos y se comparó su desempeño. Los resultados revelan que, aunque ambos modelos presentan una coherencia similar, CFDTM supera notablemente a DETM en términos de diversidad de tópicos y agrupación de documentos, lo que indica una mejor captura de la evolución temporal de los tópicos en la investigación médica.

# Abstract

Topic models are essential tools for extracting latent thematic structures from large text corpora. In this study, the Dynamic Embedding Topic Model (DETM) and the Chain-Free Dynamic Topic Model (CFDTM) were evaluated on a dataset composed of articles from the Havana Journal of Medical Sciences. The goal was to identify which model best fits the data, using metrics of topic coherence, topic diversity, clustering, and classification. Through a series of experiments, the optimal hyperparameter configurations were determined for both models, and their performance was compared. The results reveal that, although both models exhibit similar coherence, CFDTM notably outperforms DETM in terms of topic diversity and document clustering, indicating a better capture of the temporal evolution of topics in medical research.

# Índice general

<b>Introducción</b>	<b>3</b>
<b>1. Estado del Arte</b>	<b>4</b>
<b>2. Detalles de Implementación</b>	<b>10</b>
2.1. Dataset y preprocesamiento de los datos . . . . .	10
2.2. Búsqueda de Hiperparámetros Óptimos . . . . .	11
<b>3. Experimentos</b>	<b>12</b>
3.1. Aspectos clave del modelo DETM . . . . .	12
3.1.1. Descripción de hiperparámetros del modelo DETM . . . . .	12
3.2. Aspectos claves del modelo CFDTM . . . . .	13
3.2.1. Descripción de hiperparámetros del modelo CFDTM . . . . .	15
3.3. Métricas utilizadas . . . . .	16
<b>4. Resultados</b>	<b>18</b>
4.1. Modelo seleccionado . . . . .	18
4.2. Tópicos extraídos . . . . .	19
<b>Conclusiones</b>	<b>28</b>
<b>Recomendaciones</b>	<b>29</b>
<b>Bibliografía</b>	<b>30</b>
<b>Referencias</b>	<b>31</b>

# Índice de figuras

2.1. Arquitectura general de TOPMOST. Cubre los escenarios más comunes de modelado de tópicos y desacopla la carga de datos, la construcción de modelos, el entrenamiento de modelos y las evaluaciones en los ciclos de vida del modelado de tópicos. (Wu, Pan y Luu 2023) . . . . .	11
4.1. Gráficas de evolución de términos asociados al tópico 'Vigilancia epidemiológica y enfermedades infecciosas' en distintos periodos. . . . .	20
4.2. Gráficas de evolución de términos asociados al tópico 'Cáncer y enfermedades tumorales' en distintos periodos. . . . .	21
4.3. Gráficas de evolución de términos asociados al tópico 'Educación médica y formación profesional' en distintos periodos. . . . .	22
4.4. Gráficas de evolución de términos asociados al tópico 'Genética y enfermedades asociadas' en distintos periodos. . . . .	22
4.5. Gráficas de evolución de términos asociados al tópico 'Lesiones y complicaciones médicas' en distintos periodos. . . . .	23
4.6. Gráficas de evolución de términos asociados al tópico 'Salud mental y bienestar' en distintos periodos. . . . .	24
4.7. Gráficas de evolución de términos asociados al tópico 'Enfermedades neurológicas y respiratorias' en distintos periodos. . . . .	25
4.8. Gráficas de evolución de términos asociados al tópico 'Casos clínicos y diagnósticos médicos' en distintos periodos. . . . .	25
4.9. Gráficas de evolución de términos asociados al tópico 'Investigación en enfermedades renales e infecciosas' en distintos periodos. . . . .	26
4.10. Gráficas de evolución de términos asociados al tópico 'Factores de riesgo cardiovascular' en distintos periodos. . . . .	27
4.11. Gráficas de evolución de términos asociados al tópico 'Odontología y ortodoncia' en distintos periodos. . . . .	27

# Introducción

Los modelos de tópicos son herramientas matemáticas-computacionales diseñadas para extraer la estructura temática latente en una colección de documentos. Estos modelos facilitan la automatización de tareas como el análisis semántico, la recomendación de contenido y la identificación de relaciones lingüísticas, incluyendo fenómenos como la sinonimia y la polisemia. Además, su versatilidad ha permitido aplicarlos en diversas áreas, como el análisis de opiniones, la categorización automática de textos y la recuperación de información.

Un modelo de tópicos analiza un conjunto de documentos, llamado corpus, dentro del cual se identifican todas las palabras únicas, formando el vocabulario. El objetivo de los modelos de tópicos es descubrir los tópicos subyacentes en estos documentos. Cada tópico se define por un conjunto de palabras del vocabulario y, a su vez, cada documento se puede asociar a una combinación de estos tópicos.

Existen diversos enfoques, como los modelos estáticos que ofrecen una visión instantánea de los tópicos, y los modelos jerárquicos que establecen relaciones de subordinación y especialización entre tópicos permitiendo reflejar como los tópicos más generales engloban a los más específicos. Sin embargo, para analizar cómo evolucionan los tópicos a lo largo del tiempo existen los modelos de tópicos dinámicos. A diferencia de los enfoques estáticos, estos modelos rastrean cómo los tópicos surgen, se modifican o se desvanecen a lo largo del tiempo, modelando la evolución de su contenido.

El campo de los modelos de tópicos ha sido objeto de un estudio exhaustivo a nivel mundial, y Cuba no es una excepción. En el país se han desarrollado modelos con un grado considerable de eficiencia, demostrando su utilidad en diversos contextos. El problema es que, a pesar de contar con modelos de tópicos eficientes en Cuba, no es posible comprender cómo evolucionan los tópicos de investigación en revistas científicas como la Revista de Ciencias Médicas de La Habana. Esta limitación dificulta la identificación de tendencias y la toma de decisiones informadas sobre la investigación.

Dado el interés en comprender cómo evolucionan los tópicos de investigación en la Revista de Ciencias Médicas de La Habana, resulta esencial utilizar modelos que puedan capturar la dimensión temporal de los datos. Los modelos de tópicos dinámicos son especialmente adecuados para este propósito, ya que permiten identificar





tendencias y patrones a lo largo del tiempo.

A partir de esta situación la presente investigación pretende responder la siguiente pregunta científica ¿Cuál de los modelos de tópicos dinámicos mejora el desempeño para indagar en la evolución de los tópicos subyacentes en la Revista de Ciencias Médicas de La Habana?

Existen diversos modelos de tópicos dinámicos que permiten capturar la evolución temática en conjuntos de datos textuales, cada uno con características particulares que los hacen adecuados para diferentes contextos. Sin embargo, destacan especialmente el *Dynamic Embedding Topic Model* (DETM) (Adjil B. Dieng 2019) y el *Chain-Free Dynamic Topic Model* (CFDTM) (Wu, Dong et al. 2024) por sus enfoques avanzados y complementarios. Estos modelos representan paradigmas contrastantes en cómo abordan la evolución temporal y semántica de los tópicos, lo que los convierte en opciones ideales para un análisis más detallado.

Por un lado, el DETM utiliza cadenas de Markov para modelar cambios graduales y asegurar la continuidad temática a lo largo del tiempo, resultando especialmente útil en datos donde los tópicos evolucionan de forma gradual como las de ciertas áreas científicas. En cambio, el CFDTM, al eliminar la dependencia de las cadenas de Markov, permite una evolución más flexible y adaptativa de los tópicos, siendo más adecuado para datos con cambios abruptos y emergentes. Este contraste asegura que ambos modelos aborden diferentes aspectos de la dinámica temporal, proporcionando una evaluación más completa de las características del conjunto de datos. Además de estas diferencias en su enfoque temporal, ambos modelos ofrecen representaciones semánticas sofisticadas. El DETM emplea *embeddings* para capturar relaciones semánticas complejas entre palabras y tópicos, mientras que el CFDTM introduce innovaciones como *Evolution-Tracking Contrastive Learning* (ETC) y *Unassociated Word Exclusion* (UWE) para garantizar la diversidad entre los tópicos y la correspondencia de las palabras de un tópico con el momento en el tiempo en que fue extraído.

Por consiguiente, para responder la pregunta científica anteriormente planteada se define el objetivo general de este estudio: determinar cuál de los modelos de tópicos dinámicos, *Dynamic Embedded Topic Model* (DETM) (Adjil B. Dieng 2019) o *Chain-Free Dynamic Topic Model* (CFDTM) (Wu, Dong et al. 2024), ofrece el mejor desempeño para rastrear la evolución de tópicos a lo largo del tiempo en la Revista de Ciencias Médicas de La Habana.

En consecuencia, para alcanzar este objetivo principal, se plantean los siguientes objetivos específicos: en primer lugar, evaluar el rendimiento de los modelos DETM y CFDTM con diferentes configuraciones de parámetros; en segundo lugar, comparar la efectividad de ambos modelos utilizando métricas como *Topic Coherence* y *Topic Diversity*, que miden la calidad y diversidad de los tópicos identificados; posteriormente, proponer un modelo adecuado para el análisis dinámico de tópicos en el contexto

de revistas científicas cubanas, específicamente en la Revista de Ciencias Médicas de La Habana para luego analizar la evolución de los tópicos relevantes en la misma, demostrando la aplicabilidad del modelo seleccionado.

Esta tesis, desarrollada en la Facultad de Matemática y Computación de la Universidad de La Habana, se estructura en cuatro capítulos principales: Estado del Arte (Capítulo 1), Detalles de Implementación (Capítulo 2), Experimentos (Capítulo 3) y Resultados (Capítulo 4). El primer capítulo presenta una revisión exhaustiva de los trabajos previos relacionados con los modelos tópicos dinámicos, mientras que los capítulos subsiguientes detallan el proceso de implementación, los experimentos realizados y los hallazgos obtenidos.

# Capítulo 1

## Estado del Arte

Los modelos de tópicos, diseñados para descubrir los tópicos latentes en colecciones de documentos, han evolucionado significativamente en las últimas décadas. Desde los enfoques bayesianos tradicionales hasta los modelos neuronales más recientes.

Los enfoques clásicos para modelos de tópicos incluyen modelos basados en factorización de matrices no negativas, como el *Latent Semantic Analysis* (LSA) [Deerwester et al. 1988](#), que inicialmente define una matriz término-documento y luego aplica descomposición en valores singulares (SVD) a dicha matriz. Mediante la descomposición las matrices resultantes capturan la relevancia de las palabras en los tópicos, los valores singulares y la representación de los documentos en el espacio temático. Al truncar en función de un  $K$  (hiperparámetro) la matriz de valores singulares y reajustar las dimensiones de las matrices restantes se obtiene una representación reducida que ofrece proporciones de tópicos por documento y palabras por tópico. LSA fue ampliamente utilizado en recuperación de información. Sin embargo, presenta limitaciones como la selección del parámetro  $K$ , el alto consumo de recursos computacionales y la dificultad de interpretar valores negativos en sus resultados.

Para superar estas deficiencias surgió el modelo *Probabilistic Latent Semantic Analysis* (PLSA) [Hofmann 1999](#), un modelo probabilístico que asume que cada documento es una distribución multinomial de tópicos latentes. PLSA utiliza una red bayesiana para modelar las relaciones entre las variables (tópicos como latentes y palabras/documentos como observables) y aplica el algoritmo *Expectation-Maximization* (EM) (Arthur P. Dempster 1977) para optimizar los parámetros.

El modelo *Latent Dirichlet Allocation* (LDA) (D. M. Blei, Ng y Jordan 2003), desarrollado por David M. Blei en 2003, es otro enfoque probabilístico que define un tópico como una distribución sobre un vocabulario predefinido. Este modelo considera que varios tópicos pueden ocurrir simultáneamente en un documento, compartiendo todos los documentos los mismos  $K$  tópicos con proporciones distintas.

El proceso de LDA consta de tres fases principales: inicialización, reasignación

y optimización. Comenzando con la inicialización, cada palabra se asigna independientemente a un tópico. Luego durante la reasignación, se ajustan las asignaciones considerando la frecuencia de uso de cada tópico en un documento y la frecuencia con la que un tópico contiene ciertas palabras, utilizando distribuciones multinomiales generadas por Dirichlet. Finalmente, en la optimización, las reasignaciones iterativas se refinan hasta alcanzar un criterio de convergencia definido por la mínima variación en las asignaciones o un número fijo de iteraciones.

*Latente Dirichlet Allocation* (LDA) sentó las bases para modelos que exploran tanto la estructura jerárquica como la evolución temporal de los tópicos. Modelos como *Raphil* (Murphy 2012) y *Pachinko Allocation Model* (Li y McCallum 2006) representan tópicos en forma de árbol, mientras que *Topic Over Time* (Wang y McCallum 2006) captura como estos tópicos evolucionan a lo largo del tiempo. Estos modelos permiten realizar análisis más detallados y sofisticados de grandes volúmenes de texto.

Si bien LDA es preciso y ampliamente usado, presenta limitaciones al trabajar con grandes volúmenes de datos debido al aumento de variables según el tamaño del corpus y la cantidad de tópicos. Las limitaciones de los modelos de tópicos tradicionales impulsaron la investigación hacia enfoques más sofisticados. Los modelos de tópicos neuronales, basados en redes neuronales profundas, surgieron como una respuesta a estas limitaciones. Estos modelos aprovechan el poder computacional de las redes neuronales para inferir los tópicos latentes en un corpus de documentos de manera más eficiente y precisa.

Los modelos de tópicos neuronales han incorporado diversas técnicas para mejorar la calidad de la inferencia y la representatividad de los tópicos extraídos. Los modelos basados en autoencoders (AE-based NTMs) utilizan redes neuronales para mapear documentos a una representación latente y aprender distribuciones de tópicos de manera eficiente. Los modelos de tópicos con *embeddings* (NTMs with Embeddings) aprovechan representaciones vectoriales de palabras y documentos para mejorar la coherencia semántica de los tópicos aprendidos. Por otro lado, los modelos de tópicos con entrenamiento adversarial (NTMs with Adversarial Training) emplean redes generativas y discriminativas para mejorar la separación y estabilidad de los tópicos inferidos. Además, los modelos de tópicos con metadatos (NTMs with Metadata) incorporan información adicional, como autores o etiquetas, para guiar la distribución de los tópicos de manera más informada. En este contexto, los modelos de tópicos dinámicos extienden estas ideas al dominio temporal, permitiendo capturar la evolución de los tópicos a lo largo del tiempo y proporcionando una visión más completa de la variabilidad en colecciones de documentos.

Los Modelos de Tópicos Neuronales Dinámicos, DNTM por sus siglas en inglés, han evolucionado principalmente en la última década. Las variantes propuestas han alcanzado gran precisión y entre ellas difieren incluso en la modelación del problema. A diferencia del enfoque tradicional donde los documentos están representados por

una bolsa de palabras <sup>1</sup> y los tópicos están representados por un vector que indica la proporción de las palabras en dicho tópico, surgieron modelos como *Dynamic Embedding Topic Model (D-ETM)* (Adj B. Dieng 2019) que sigue un enfoque diferente representando palabras y tópicos como *embeddings* <sup>2</sup>. Este modelo tiene como bases a los modelos LDA (D. M. Blei, Ng y Jordan 2003), *Dynamic LDA (D-LDA)* (Blei y Lafferty 2006) y *Embeddings Topic Model (ETM)* (Dieng et al. 2019). D-LDA es una extensión de LDA que utiliza una serie de tiempo probabilística para permitir que los tópicos varíen suavemente a lo largo del tiempo, pero disminuye su factibilidad al presentar las mismas limitaciones que LDA. Para sobrepasar esas limitaciones surge ETM que representa los tópicos como un vector en el espacio de los *embeddings* de las palabras para luego usar el producto punto entre los *embeddings* de las palabras y los *embeddings* de los tópicos para definir la distribución de palabras por tópico. De manera similar a D-LDA, el D-ETM involucra una serie de tiempo probabilística para permitir que los tópicos varíen suavemente a lo largo del tiempo. Sin embargo, cada tópico en el D-ETM es un vector que varía con el tiempo en el espacio en el que fue proyectado. Al igual que en el ETM, la probabilidad de cada palabra bajo el D-ETM es una distribución categórica cuyo parámetro natural depende del producto interno entre el *embedding* de la palabra y el *embedding* de su tópico asignado. En contraste con el ETM, los *embedding* de tópico del D-ETM varían con el tiempo y son aproximados mediante el mecanismo de inferencia variacional (Jordan et al. 1999; D. M. Blei, Kucukelbir y McAuliffe 2017). Para escalar el algoritmo a grandes conjuntos de datos en este modelos fueron utilizados el submuestreo de datos (Hoffman et al. 2013) y la amortización (Gershman y Goodman 2014) con un parámetro de aproximación variacional estructurado parametrizado por una red neuronal de memoria a corto y largo plazo (LSTM) (Hochreiter y Schmidhuber 1997).

Surgieron luego, en 2022 los modelos NetDTM y NetDTM+ (Zhang y Lauw 2022, 17–23 Jul) que funcionan bajo el supuesto de que la naturaleza temporal se relaciona no solo con el momento en que se crea un documento, sino también con cómo los documentos creados en diferentes momentos pueden formar vínculos. Dado que una red de documentos no emerge en su totalidad de forma imprevista, los creadores afirman que teniendo una red inicial a medida que pasa el tiempo crece no solo el corpus sino también la conectividad de la red. Estos modelos aprovechan esta estructura en red al incorporar las conexiones entre documentos como una fuente adicional de información, mejorando así la capacidad del modelo para capturar patrones temporales y relacionales en los datos.

Otro enfoque novedoso fue el usado en el modelo *Aligned Neural Topic Model*

<sup>1</sup>En procesamiento de lenguaje natural, la bolsa de palabras es una representación simplificada de texto donde un documento se describe mediante la frecuencia de aparición de cada palabra, ignorando el orden y la gramática.

<sup>2</sup>Un *embedding* es el mapeo de una variable categórica a un vector numérico.

(ANTM) (Rahimi et al. 2023) que descubre la evolución de los tópicos usando *clusters*. ANTM descubre tópicos en evolución a través de un algoritmo de ventanas deslizantes superpuestas para la agrupación temporal de documentos. El algoritmo utilizado, llamado *Aligned Clustering*, consiste en segmentar archivos en segmentos superpuestos, realizar agrupación secuencial basada en densidad y alinear *clusters* de documentos similares adyacentes a lo largo de diferentes períodos. El *Aligned Clustering* permite la identificación de conjuntos de tópicos similares que abarcan múltiples períodos de tiempo y, sin embargo, son lo suficientemente diferentes como para mostrar algún tipo de evolución. ANTM aprovecha los grandes modelos de lenguaje (LLMs) preentrenados, como Data2Vec (Baeovski et al. 2022), que predice representaciones latentes de los documentos de manera auto-supervisada utilizando una arquitectura *transformer* estándar.

Siguiendo la idea de establecer conexiones entre los documentos del corpus ubicados en distintos espacios de tiempo surge el modelo *Dynamic Structured Neural Topic Model* (DSNTM) (Miyamoto et al. 2023, julio), un modelo que calcula las dependencias entre los tópicos a través del tiempo. DSNTM modela dichas dependencias basándose en un mecanismo de autoatención (Vaswani et al. 2023; Lin et al. 2017), que al observar sus pesos revela como los tópicos pasados se ramifican o fusionan en nuevos tópicos siendo posible predecir tópicos emergentes y evaluar cuantitativamente qué tópicos pasados contribuyen a la aparición de nuevos tópicos. Este modelo introduce la regularización de citas que lleva a los pesos del mecanismo de atención a reflejar las relaciones de citación entre documentos y como resultado es posible modelar tanto el texto como las citas conjuntamente, mejorando la calidad de los tópicos inferidos y capturando con precisión sus transiciones. Debido al alto poder expresivo del mecanismo de autoatención y la información adicional de citas DSNTM resalta en factibilidad a la hora de detectar complejos procesos de ramificación y fusión de tópicos a lo largo del tiempo.

En ocasiones categorizar un **modelo tópico** puede ser algo ambiguo. Un ejemplo de esto es el modelo *Neural Dynamic Focused Topic Model* (NDF-TM) (Cvejowski et al. 2023), un modelo que en lugar de modelar la evolución de los tópicos modela las actividades de estos a lo largo del tiempo. Cabe destacar que las actividades de los tópicos evolucionan con el tiempo, pero sus tópicos por sí mismos son invariantes. Por lo tanto, este método no se adhiere con precisión a la definición original del modelado tópico dinámico (Wu, Nguyen y Luu 2024). En este enfoque se busca desacoplar la probabilidad de que un tópico esté activo de su proporción mediante la introducción de secuencias de variables aleatorias de Bernoulli, que seleccionan los tópicos activos para un documento dado en un instante particular de tiempo. Para lograr el **desacoplamiento** sigue una lógica similar a la de los priors no paramétricos, como el proceso de buffet indio sobre matrices binarias infinitas, tanto en entornos estáticos (Williamson et al. 2010) como dinámicos (Perrone et al. 2016), pero aprovecha el truco de

reparametrización para realizar inferencia variacional estocástica (Kingma y Welling 2022). El resultado es un modelo escalable que permite que el número instantáneo de tópicos activos por documento fluctúe y que desacopla explícitamente la proporción de tópicos de su actividad, ofreciendo así capas novedosas de interpretabilidad y transparencia en la evolución de los tópicos a lo largo del tiempo.

En su mayoría los modelos existentes hasta el momento relacionaban los tópicos a través de cadenas de Markov para capturar su evolución y disminuían su efectividad al extraer tópicos repetidos<sup>3</sup> y ubicar tópicos en espacios de tiempo incorrectos.

El uso de cadenas de Markov para encadenar tópicos tiende a agruparlos, lo que dificulta su diferenciación y limita su capacidad para capturar plenamente la semántica de sus respectivos períodos de tiempo, según se plantea en Wu, Dong et al. 2024. Además, este enfoque puede forzar una relación excesiva entre los tópicos a lo largo de diferentes períodos, reduciendo su asociación con los segmentos de tiempo específicos y ocultando los tópicos genuinos en esos segmentos (Wu, Dong et al. 2024). Para resolver los problemas anteriores surge el modelo *Chain-Free Dynamic Topic Model* (CFDTM) (Wu, Dong et al. 2024) que rompe la tradición de encadenar tópicos a través de cadenas de Markov y propone un nuevo enfoque respaldado por dos submodelos: *Evolution-Tracking Contrastive Learning* (ETC) y *Unassociated Word Exclusion* (UWE). ETC construye adaptativamente relaciones positivas y negativas entre los tópicos dinámicos. En este contexto  $\varphi_{k'}^{(t)}$  representa al tópico  $k$  en el momento  $t$ . Para rastrear la evolución de los tópicos ETC construye relaciones positivas para los pares  $(\varphi_k^{(t-1)}, \varphi_k^{(t)})$  y más importante aún, construye relaciones negativas entre los pares  $(\varphi_k^{(t)}, \varphi_{k'}^{(t)})$  con  $k \neq k'$ . Estas relaciones negativas alientan a los tópicos dentro de un mismo espacio de tiempo a ser distintos, manteniendo la diversidad de tópicos y así aliviando el problema de los tópicos repetitivos. Por otro lado UWE encuentra las palabras más relevantes de los tópicos en cada segmento de tiempo e identifica cuáles de ellas no pertenecen a este segmento de tiempo como palabras no asociadas, así excluye explícitamente estas palabras no asociadas de los tópicos para refinar la semántica de los mismos. Siguiendo la práctica común (Blei y Lafferty 2006; Adji B. Dieng 2019; Miyamoto et al. 2023, julio), el modelo usa el hiperparámetro  $\lambda(t)$  para ajustar adaptativamente las intensidades de evolución entre segmentos de tiempo. Si los tópicos evolucionan ligeramente entre los segmentos de tiempo  $t-1$  y  $t$ , se utiliza un  $\lambda(t)$  grande; de lo contrario, un  $\lambda(t)$  pequeño si evolucionan dramáticamente. El proceso generativo en CFDTM usa un VAE donde el *encoder* es reutilizado para documentos en diferentes espacios de tiempo para garantizar la eficiencia de parámetros. Como resultado de los métodos novedosos empleados, CFDTM alcanzó el mejor índice de TC<sup>4</sup> con mejoras significantes respecto al resto de los modelos

<sup>3</sup>Los tópicos dentro de un segmento de tiempo se consediran repetitivos al presentar una semántica similar.

<sup>4</sup>TC: Topic Coherence. Métrica usada para medir la coherencia entre las palabras asignadas a

existentes.

En conclusión, los Modelos de Tópicos Neuronales Dinámicos han evolucionado considerablemente, incorporando técnicas avanzadas como *embeddings* y mecanismos de autoatención para superar las limitaciones de los enfoques tradicionales. Las innovaciones recientes han enriquecido la capacidad de capturar la evolución semántica y temporal de los tópicos. Estos avances no solo mejoran la interpretabilidad y la diversidad de los tópicos inferidos, sino que también abren nuevas posibilidades para abordar desafíos complejos en el análisis de datos textuales dinámicos.





# Capítulo 2

## Detalles de Implementación



### 2.1. Dataset y preprocesamiento de los datos

El conjunto de datos empleado en esta investigación fue compilado a partir de los artículos pertenecientes a la Revista de Ciencias Médicas de La Habana. Dicha revista científica cuenta con una colección de más de 1500 artículos disponibles en su sitio web oficial, los cuales constituyeron la fuente para la extracción del dataset.

Una vez adquirido el corpus, los documentos fueron sometidos a una fase de procesamiento que sigue la metodología implementada en el módulo TopMost (Wu, Pan y Luu 2023) del lenguaje Python, la cual abarca una serie de etapas cruciales para preparar el texto para el modelado de tópicos. Inicialmente, se lleva a cabo la limpieza del texto, que incluye la eliminación de etiquetas HTML, el cambio a minúsculas, la supresión de direcciones de correo electrónico y menciones a usuarios, y además la normalización de espacios en blanco.

Posteriormente, se realiza la tokenización, donde el texto se segmenta en unidades individuales o *tokens*. A continuación, se filtran estos *tokens*, eliminando las palabras de alta frecuencia pero de bajo contenido informativo (*stopwords*) y aquellos *tokens* que no cumplen con ciertos criterios, como la presencia exclusiva de números o una combinación de letras y números. Una vez obtenidos los *tokens* relevantes, se construye el vocabulario, formado por el conjunto de todos los *tokens* presentes en el corpus.

Finalmente, los documentos son representados como bolsas de palabras (*Bag-of-Words*), donde cada documento se convierte en un vector que indica la frecuencia de cada término del vocabulario en dicho documento. Para enriquecer la representación semántica de los términos, se generan *embeddings* para lo que se empleó la versión en español del modelo pre-entrenado de *embeddings FastText* (Bojanowski et al. 2016). La elección de este modelo se basó en su capacidad para capturar relaciones semánticas entre palabras, incluso para aquellas menos frecuentes, gracias a su enfoque en subpalabras (*subword information*).

## 2.2. Búsqueda de Hiperparámetros Óptimos

Para el entrenamiento de los modelos de tópicos dinámicos DETM y CFDTM y su posterior evaluación, se utilizó el módulo de Python TopMost<sup>1</sup> (Wu, Pan y Luu 2023) que usa PyTorch (Paszke et al. 2019) como marco de trabajo de redes neuronales para modelos de tópicos neuronales. La elección de TopMost como marco de implementación se fundamenta en su facilidad de uso, su eficiencia computacional y su capacidad para trabajar con modelos de tópicos de última generación como DETM y CFDTM. Esta herramienta de código abierto proporciona una interfaz eficiente y flexible para la evaluación de los modelos y el preprocesamiento de los datos, lo que facilita el desarrollo y la comparación de diferentes enfoques de modelado de tópicos. La arquitectura de TopMost se muestra en la Figura 2.1.

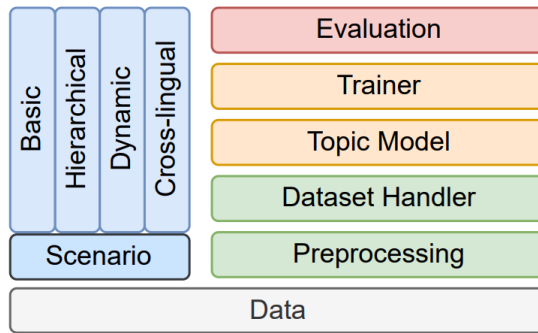


Figura 2.1: Arquitectura general de TOPMOST. Cubre los escenarios más comunes de modelado de tópicos y desacopla la carga de datos, la construcción de modelos, el entrenamiento de modelos y las evaluaciones en los ciclos de vida del modelado de tópicos. (Wu, Pan y Luu 2023)

El entrenamiento de los modelos de tópicos dinámicos DETM y CFDTM se llevó a cabo utilizando la plataforma Kaggle, un entorno basado en la nube que proporciona recursos computacionales para el desarrollo de proyectos de aprendizaje automático. Kaggle ofrece acceso a diferentes tipos de *hardware*, incluyendo CPU y GPU, lo que permite acelerar el proceso de entrenamiento de modelos complejos. Cada entorno de ejecución en Kaggle dispone de hasta 16 GB de VRAM en el caso de las GPU y de hasta 30 GB de RAM en el caso de las CPU.

<sup>1</sup><https://github.com/bobxwu/topmost>



## Capítulo 3

# Experimentos

La eficiencia de los modelos de tópicos dinámicos depende en gran medida del dataset empleado, en este caso los artículos pertenecientes a la Revista de Ciencias Médicas de La Habana, ya que las características específicas del corpus influyen directamente en la capacidad del modelo para descubrir patrones significativos. Se llevó a cabo una exploración exhaustiva de las configuraciones de hiperparámetros para dos modelos: CFDTM y DETM, con el objetivo de identificar aquellos valores que maximizan la coherencia de los tópicos, la diversidad y el rendimiento general en este corpus.

### 3.1. Aspectos clave del modelo DETM

DETM aborda la evolución temporal modelando los tópicos como secuencias de *embeddings* en un espacio continuo. La implementación de DETM en TopMost, utiliza redes neuronales recurrentes (RNN), específicamente LSTM (*Long Short-Term Memory*), para capturar la dinámica temporal de los tópicos. Los tópicos se representan como puntos en el espacio de *embeddings*, y la LSTM modela la evolución de estos puntos a lo largo del tiempo. La distribución de tópicos de cada documento se representa como una combinación de estos *embeddings* de tópicos, y el modelo aprende la evolución de estos *embeddings* en el tiempo. El proceso de inferencia se basa en la aproximación variacional, y se utilizan redes neuronales *feedforward* para parametrizar las distribuciones variacionales.

#### 3.1.1. Descripción de hiperparámetros del modelo DETM

- **Número de tópicos:** Referente al número de tópicos latentes que el modelo intentará descubrir en el corpus. Un valor muy bajo puede mezclar tópicos distintos mientras que un valor muy alto puede granular un mismo tópico.

- **Dimensión de  $\rho$ :** Define la dimensión del espacio de *embeddings* ( $\rho$ ) donde se encuentran las palabras.  $\rho$  corresponde a las representaciones de palabras en el modelo.
- **Tamaño del espacio oculto de  $q(\theta)$ :** Define la cantidad de unidades ocultas en la red neuronal que modela la distribución variacional de los documentos.
- **Función de activación para  $\theta$ :** Especifica la función de activación usada en la red neuronal para modelar  $\theta$ , que representa las distribuciones de tópicos por documento.
- **Número de capas para  $\eta$ :** Define el número de capas en la red recurrente que modela la evolución de los tópicos a lo largo del tiempo.
- **Tamaño del espacio oculto en  $\eta$ :** Define la cantidad de unidades ocultas en la RNN que modela la dinámica temporal de los tópicos.
- **Varianza del prior de  $\eta$ :** Especifica la varianza de la distribución previa sobre los cambios en los tópicos ( $\eta$ ). .
- **Tasa de Aprendizaje:** Controla el tamaño de los pasos que da el optimizador durante el entrenamiento. Una tasa de aprendizaje alta puede llevar a la inestabilidad, mientras que una muy baja puede resultar en una convergencia lenta.
- **Dropout en el encoder:** Define la tasa de dropout aplicada a la red neuronal que codifica la distribución  $q(\theta)$ , usada para estimar la distribución de tópicos por documento.
- **Dropout en la RNN para  $\eta$ :** Define la tasa de dropout aplicada en la RNN que modela la evolución temporal de los tópicos ( $\eta$ ).
- **Número de épocas:** Define el número de veces que el modelo recorrerá el conjunto de datos de entrenamiento completo.

El espacio de búsqueda para encontrar la configuración óptima de hiperparámetros se muestra en las tablas 3.1 y 3.2.

## 3.2. Aspectos claves del modelo CFDTM

CFDTM es un modelo de tópicos dinámico que, a diferencia de los enfoques tradicionales basados en cadenas de Markov, introduce un nuevo método de aprendizaje

Tabla 3.1: Espacio de búsqueda de hiperparámetros del modelo

Hiperparámetro	Valores
num_topics	{20, 50, 100}
rho_size	{200, 300, 400}
en_units	{400, 800, 1200}
eta_nlayers	{2, 3, 4}
delta	{0.001, 0.005, 0.01, 0.05}
theta_act	{anh, softplus, relu, rrelu, leakyrelu, elu, selu, glu}
eta_hidden_size	{100, 200, 300}

Tabla 3.2: Espacio de búsqueda de hiperparámetros de entrenamiento

Hiperparámetro	Valores
enc_drop	{0.0, 0.2, 0.3, 0.5}
eta_dropout	{0.0, 0.2, 0.3, 0.5}
learning_rate	{1e-3, 5e-4, 0.02}
epochs	{200, 400, 800}

contrastivo con seguimiento de la evolución que construye relaciones de similitud entre los tópicos dinámicos para modelar la evolución de los tópicos en el tiempo. Esto no solo permite rastrear la evolución de los tópicos, sino que también mantiene la diversidad de los mismos, mitigando el problema de los tópicos repetitivos. De la misma forma para evitar tópicos no asociados, implementa además un método de exclusión de palabras no asociadas que elimina consistentemente palabras no relacionadas de los tópicos descubiertos.

En términos de arquitectura, CFDTM utiliza *embeddings* de palabras pre-entrenados como entrada. Estos *embeddings* se proyectan a un espacio latente común mediante una capa de proyección lineal. La evolución de los tópicos se modela a través de la proyección de los *embeddings* de los tópicos en este espacio latente. El aprendizaje contrastivo se implementa utilizando la pérdida InfoNCE, que compara las similitudes entre los *embeddings* de los tópicos en diferentes instantes de tiempo. Para la exclusión de palabras no asociadas, se utiliza una red neuronal *feedforward* que toma como entrada el *embedding* de la palabra y el *embedding* del tópico, y produce una puntuación de asociación.

### Aprendizaje Contrastivo de Seguimiento de Evolución (Evolution-Tracking Contrastive Learning)

En lugar de depender de una cadena de Markov, CFDTM utiliza un enfoque de aprendizaje contrastivo para alinear los *embeddings* de los tópicos en diferentes

instantes de tiempo. Este alienta a que los *embeddings* de tópicos del mismo tópico en instantes de tiempo cercanos sean similares, mientras que los *embeddings* de tópicos diferentes sean distinguibles. Esto permite que el modelo capture la evolución gradual de los tópicos sin imponer una estructura de dependencia estricta.

### Exclusión de Palabras No Asociadas (Unassociated Word Exclusion)

Para mejorar la coherencia de los tópicos, CFDTM introduce un mecanismo para excluir palabras que no están fuertemente asociadas con ningún tópico en un instante de tiempo dado. Esto se logra mediante una red neuronal *feedforward* que predice la probabilidad de que una palabra pertenezca a un tópico específico. Las palabras con baja probabilidad son excluidas del cálculo de la pérdida, lo que ayuda a refinar los tópicos y a eliminar el ruido.

#### 3.2.1. Descripción de hiperparámetros del modelo CFDTM

- **Número de tópicos:** Referente al número de tópicos latentes que el modelo intentará descubrir en el corpus. Un valor muy bajo puede mezclar tópicos distintos, mientras que un valor muy alto puede granular un mismo tópico.
- **Unidades en la capa encoder:** Número de neuronas en la primera capa del encoder.
- **Dropout en la red:** Define la tasa de *dropout* aplicada para regularizar el modelo.
- **Temperatura de  $\beta$ :** Factor de escala para la distribución  $\beta$  de los tópicos en el modelo.
- **Temperatura en contraste:** Controla la suavidad en la función de contraste utilizada en el aprendizaje de *embeddings*.
- **Peso para palabras negativas:** Ponderación para las palabras negativas en el mecanismo de contraste.
- **Peso para palabras positivas:** Ponderación para palabras positivas en la función de contraste.
- **Peso para *Unassociated Word Exclusion* (UWE):** Controla la penalización por palabras que no están asociadas a ningún tópico en el momento  $t$ .
- **Número de palabras negativas  $K$ :** Define cuántas palabras negativas se consideran en la pérdida contrastiva.

- **Tasa de aprendizaje:** Controla la magnitud de los cambios en los parámetros del modelo durante el entrenamiento.
- **Tamaño del *batch*:** Define cuántos ejemplos de entrenamiento se procesan simultáneamente antes de actualizar los parámetros del modelo.
- **Número de épocas:** Cantidad de veces que el modelo recorrerá el conjunto de datos completo.

El espacio de búsqueda para encontrar la configuración óptima de hiperparámetros se muestra en las tablas 3.3 y 3.4.

Tabla 3.3: Espacio de búsqueda de hiperparámetros del modelo

Hiperparámetro	Valores
num_topics	{20, 50, 100}
enl_units	{50, 100, 200, 300}
dropout	{0.0, 0.1, 0.3, 0.5}
beta_temp	{0.5, 1.0, 1.5, 2.0}
temperature	{0.05, 0.1, 0.2, 0.5}
weight_neg	{1e6, 5e6, 1e7, 1e8}
weight_pos	{1.0, 10.0, 100.0}
weight_UWE	{1e2, 1e3, 1e4}
neg_topk	{5, 10, 15, 20, 30}

Tabla 3.4: Espacio de búsqueda de hiperparámetros de entrenamiento

Hiperparámetro	Valores
learning_rate	{0.001, 0.002, 0.005, 0.01}
batch_size	{100, 200, 300, 500}
num_epoch	{400, 600, 800, 1000}

### 3.3. Métricas utilizadas

Para evaluar la calidad del modelo de tópicos, se utilizaron principalmente dos métricas: Coherencia del Tópico (TC - *Topic Coherence*) y Diversidad del Tópico (TD - *Topic Diversity*).

### Coherencia del Tópico (TC)

La coherencia del tópico mide qué tan relacionadas están las palabras más representativas de un tópico. Esta métrica evalúa si las palabras principales de un tópico aparecen juntas con frecuencia en el corpus de documentos. Para evaluar esta coherencia, se utilizó la métrica CV (Röder et al. 2015). Una mayor coherencia indica que el tópico es más interpretable y significativo.

### Diversidad del Tópico (TD)

La diversidad del tópico mide qué tan distintos son los tópicos descubiertos. Se calcula como la proporción de palabras principales de un tópico que no se repiten en otros tópicos. Un alto valor de TD indica que los tópicos son bien diferenciados entre sí, evitando redundancias y asegurando que el modelo capture una variedad más amplia de tópicos.

Además de estas métricas principales, se evaluó la calidad de las distribuciones de tópicos en los documentos mediante dos tareas extrínsecas: clasificación de textos y *clustering* de documentos.

### Clasificación de Textos

Para la clasificación de textos se entrena un clasificador tradicional (SVM), utilizando las distribuciones documento-tópico como características. Luego, se predicen las etiquetas de otros documentos, permitiendo evaluar qué tan bien el modelo organiza los documentos en función de sus tópicos.

### *Clustering* de Documentos

Para el *clustering* en este caso, se asigna a cada documento el tópico más significativo en su distribución documento-tópico y se analiza qué tan bien se agrupan los documentos dentro de estos tópicos.

Si bien las métricas de *clustering* y clasificación nos permiten evaluar la calidad del modelo en términos de su capacidad para organizar documentos, no fueron utilizadas para la selección del modelo óptimo, sino solo como una referencia adicional para analizar la calidad de los tópicos finales generados.



# Capítulo 4

## Resultados

### 4.1. Modelo seleccionado

Para determinar el modelo que mejor se adapta al conjunto de datos, se realizaron múltiples experimentos con los modelos de tópicos dinámicos DETM y CFDTM sobre un corpus compuesto por artículos extraídos de la Revista de Ciencias Médicas de La Habana. Se probaron diversas configuraciones de hiperparámetros y, tras cientos de experimentos, se encontraron las configuraciones óptimas que produjeron los mejores resultados. Dichas configuraciones se muestran en las tablas 4.1 y 4.2.

Tabla 4.1: Configuración óptima de hiperparámetros para DETM.

Hiperparámetro	Valor
Número de tópicos ( <code>num_topics</code> )	20
Tamaño de $\rho$ ( <code>rho_size</code> )	400
Unidades en encoder ( <code>en_units</code> )	1200
Tamaño oculto de $\eta$ ( <code>eta_hidden_size</code> )	300
Capas en $\eta$ ( <code>eta_nlayers</code> )	4
Dropout en encoder ( <code>enc_drop</code> )	0.2
Dropout en $\eta$ ( <code>eta_dropout</code> )	0.0
Delta ( <code>delta</code> )	0.001
Activación en $\theta$ ( <code>theta_act</code> )	SELU
Tasa de aprendizaje ( <code>learning_rate</code> )	0.0005
Épocas ( <code>epochs</code> )	400

Los resultados cuantitativos obtenidos en función de las métricas propuestas en el Capítulo 3 se presentan en la Tabla 4.3.

Si bien la coherencia de los tópicos extraídos por ambos modelos es similar, se observa una diferencia notable en la diversidad de los tópicos generados. En particu-

Tabla 4.2: Configuración óptima de hiperparámetros para CFDTM.

Hiperparámetro	Valor
Número de tópicos ( <code>num_topics</code> )	20
Unidades en capa 1 ( <code>en1_units</code> )	50
Dropout ( <code>dropout</code> )	0.1
Temperatura beta ( <code>beta_temp</code> )	1.5
Temperatura ( <code>temperature</code> )	0.05
Peso negativo ( <code>weight_neg</code> )	5,000,000
Peso positivo ( <code>weight_pos</code> )	100
Peso UWE ( <code>weight_UWE</code> )	1,000
Top-k negativo ( <code>neg_topk</code> )	20
Tasa de aprendizaje ( <code>learning_rate</code> )	0.001
Tamaño de lote ( <code>batch_size</code> )	200
Épocas ( <code>num_epoch</code> )	800

Tabla 4.3: Resultados cuantitativos de los modelos DETM y CFDTM.

Métrica	DETM	CFDTM
TC ( <code>dynamic_TC</code> )	0.4587	0.5102
TD ( <code>dynamic_TD</code> )	0.6520	0.9455
Pureza de Clustering ( <code>Purity</code> )	0.2888	0.4210
Información Mutua Normalizada ( <code>NMI</code> )	0.1481	0.3044
Exactitud en Clasificación ( <code>acc</code> )	0.3252	0.4498
Macro-F1	0.1114	0.2725

lar, CFDTM muestra una diversidad significativamente mayor en comparación con DETM, lo que sugiere que este modelo captura de manera más efectiva los cambios drásticos en los tópicos latentes a lo largo del tiempo en los artículos de la Revista de Ciencias Médicas de La Habana. Estos resultados indican que CFDTM extrae tópicos con mayor calidad y representa mejor la evolución temática de la revista en diferentes periodos.

## 4.2. Tópicos extraídos

Para facilitar la interpretación de los resultados, la evolución de cada tópico se presenta mediante gráficas que permiten visualizar los cambios en la relevancia de los términos claves de un tópico a lo largo del tiempo de manera clara e intuitiva.

Cada gráfico asociado a un tópico muestra la variación del peso de ciertas palabras claves en distintos periodos, lo que posibilita el análisis de cómo han evolucionado

los tópicos a lo largo de los años. Aunque estas palabras no siempre son las más representativas del tópico, su estudio ofrece una perspectiva sobre las tendencias y transformaciones en el contenido temático.

Para cada tópico, se presentan cuatro gráficas correspondientes a los periodos 2003-2008, 2009-2014, 2015-2019 y 2020-2024, organizadas en orden ascendente. En cada gráfica se representa la probabilidad de las palabras de pertenecer al tópico en cuestión. Además, a excepción del primer periodo (2003-2008), en las gráficas se muestran en color gris las palabras pertenecientes al periodo anterior, mientras que aquellas que se mantienen o emergen en el nuevo periodo se resaltan con distintos colores, lo que permite identificar con mayor claridad su incorporación y evolución dentro del tópico analizado.

### Vigilancia epidemiológica y enfermedades infecciosas (*topic#1*)

En los primeros años (2003-2010), se observa un énfasis en la vigilancia sanitaria, publicaciones científicas y modelos de productividad en instituciones de salud. A partir de 2011, surgen menciones a Cuba y Brasil con términos como intoxicaciones y medidas sanitarias. Entre 2018 y 2020, la llegada de COVID-19 domina los registros, con referencias a tasas de infección, medidas de control y la OMS. Desde 2021 hasta 2024, aparecen términos como telemedicina, seguridad digital, vacunas y hospitales, reflejando una evolución hacia modelos digitales y de investigación en salud. (Figura 4.1)

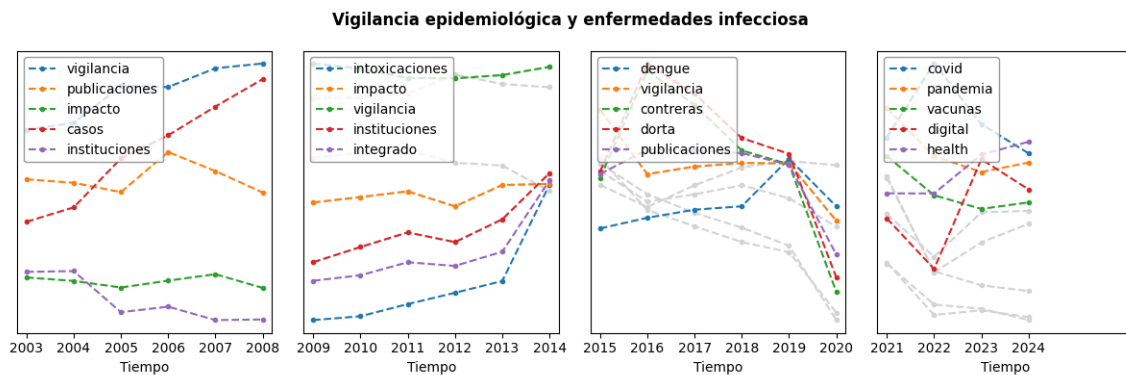


Figura 4.1: Gráficas de evolución de términos asociados al tópico 'Vigilancia epidemiológica y enfermedades infecciosas' en distintos periodos.

### Cáncer y enfermedades tumorales (*topic#3*)

Inicialmente, entre 2003 y 2008, el enfoque principal se centraba en el diagnóstico y las variantes de cáncer, abarcando términos como tumores, tuberculosis, virus y

literatura médica. Posteriormente, de 2009 a 2014, los estudios se volvieron más específicos, mencionando biopsias y neoplasias en áreas como pulmón, piel y cabeza. Un avance significativo se observa entre 2015 y 2020 con la aparición de términos como carcinoma, imaging y contraste, reflejando el desarrollo de técnicas de diagnóstico por imágenes. Finalmente, en el período de 2021 a 2024, se incorporan términos como patogenia, laringe, suero, pleura y padecimiento, sugiriendo un enfoque más profundo en la biología del cáncer. (Figura 4.2)

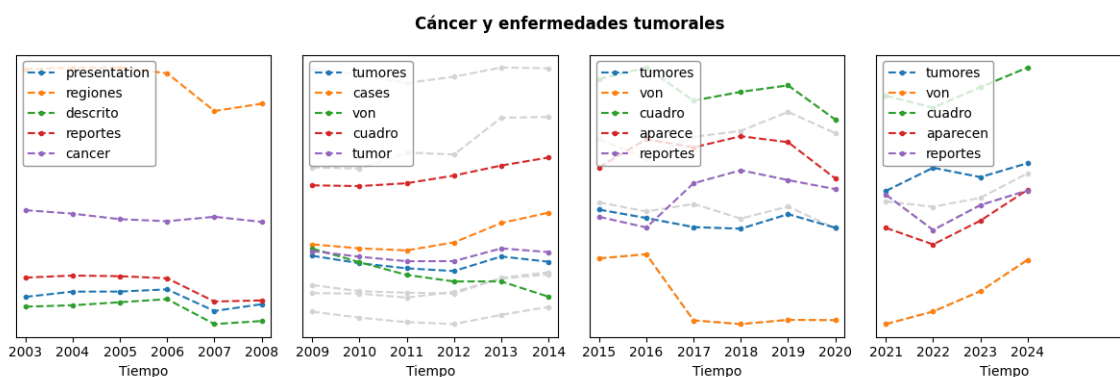


Figura 4.2: Gráficas de evolución de términos asociados al tópico 'Cáncer y enfermedades tumorales' en distintos periodos.

#### Educación médica y formación profesional (*topic#4*)

La evolución del enfoque en el aprendizaje y la formación profesional se evidencia a través de distintos periodos. Inicialmente, entre 2003 y 2008, se priorizaba el aprendizaje y la formación profesional, destacando términos como estrategias, docentes, programas, especialidad y asignaturas. Luego, de 2009 a 2014, se observa un interés en la modernización de la enseñanza con la aparición de palabras como *software*, educativo y pertinencia curricular. Entre 2015 y 2020, se integran conceptos como competencias, educación digital y educación en salud, lo que sugiere un cambio hacia un modelo más dinámico. Finalmente, en el período de 2021 a 2024, se menciona el uso de videos y estrategias digitales para la formación de médicos residentes, resaltando la importancia de las herramientas digitales en la formación actual. (Figura 4.3)

#### Genética y enfermedades asociadas (*topic#6*)

Primeramente, entre 2003 y 2008, el enfoque principal se centraba en la relación entre genes y enfermedades de la piel, explorando conceptos como carcinoma,

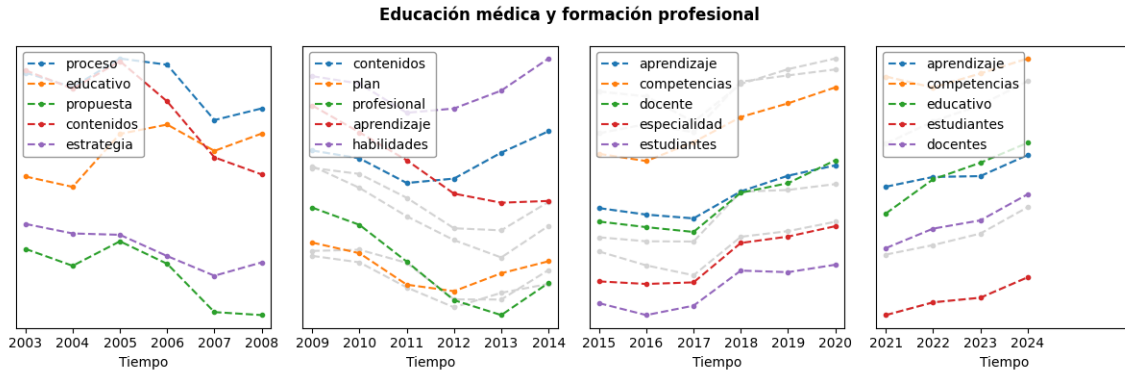


Figura 4.3: Gráficas de evolución de términos asociados al tópico 'Educación médica y formación profesional' en distintos periodos.

receptores y mutaciones. Posteriormente, de 2009 a 2014, se amplió el espectro de investigación para incluir estudios sobre el *Helicobacter pylori* y la neoplasia gástrica, lo que sugiere un interés creciente en la relación entre diferentes factores y enfermedades. Un avance notable se observa entre 2015 y 2020 con la integración de palabras como inmunología, hormonas y enfermedades raras, lo que refleja una comprensión más profunda de la complejidad de los sistópicos biológicos. Finalmente, en el período de 2021 a 2024, la aparición de conceptos como alteraciones genéticas y patología molecular indica un enfoque cada vez mayor en la genética aplicada a la medicina, lo que promete avances significativos en el diagnóstico y tratamiento de enfermedades. (Figura 4.4)

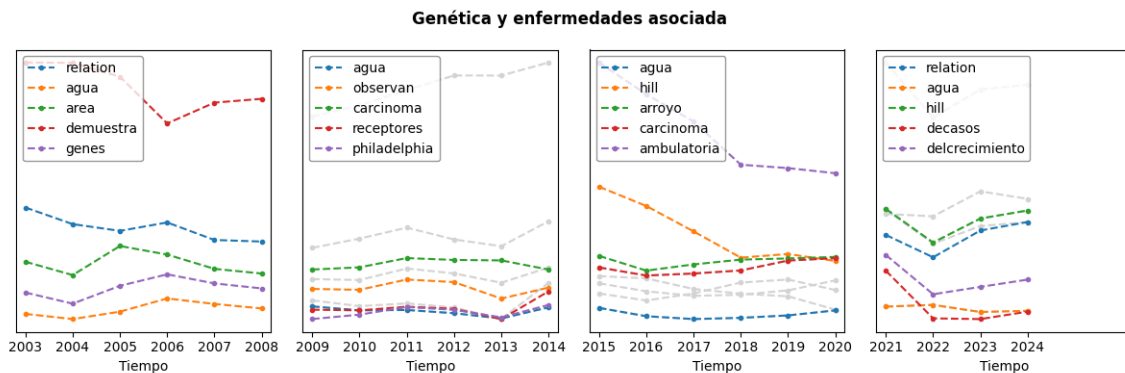


Figura 4.4: Gráficas de evolución de términos asociados al tópico 'Genética y enfermedades asociadas' en distintos periodos.

### Lesiones y complicaciones médicas (*topic#7*)

Inicialmente, entre 2003 y 2008, el enfoque principal se centraba en urgencias médicas y fracturas, abarcando conceptos como accidentes, mama, estadio y bucal. Posteriormente, de 2009 a 2014, se amplió el espectro de investigación para incluir fracturas maxilofaciales y lesiones temporomandibulares, lo que sugiere un interés creciente en la relación entre diferentes tipos de lesiones y su tratamiento. Un avance significativo se observa entre 2015 y 2020 con la integración de conceptos como cirugía reconstructiva y melanoma, lo que refleja una comprensión más profunda de la complejidad de los procedimientos quirúrgicos y las enfermedades de la piel. Finalmente, en el período de 2021 a 2024, la aparición de conceptos como anastomosis, reflujo y complicaciones postoperatorias indica un enfoque cada vez mayor en la cirugía avanzada, lo que promete avances significativos en el tratamiento de diversas condiciones médicas. (Figura 4.5)

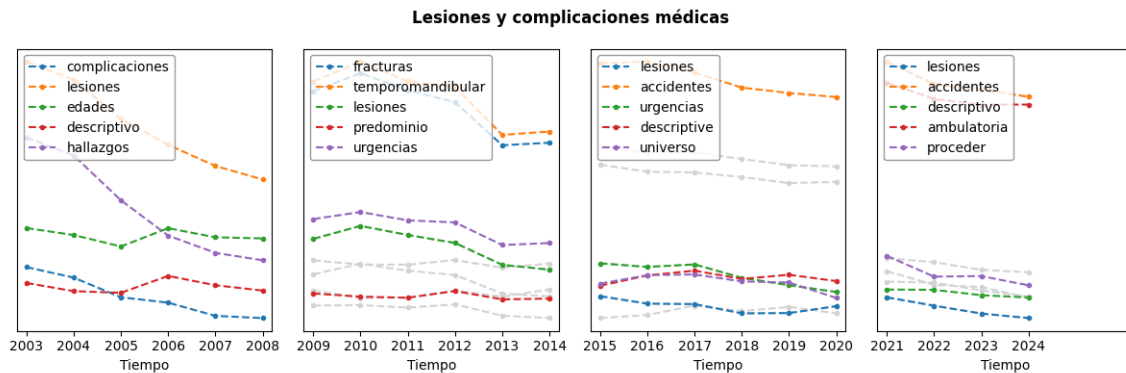


Figura 4.5: Gráficas de evolución de términos asociados al tópico 'Lesiones y complicaciones médicas' en distintos periodos.

### Salud mental y bienestar (*topic#9*)

Al comienzo, entre 2003 y 2008, el enfoque principal se centraba en el bienestar y la calidad de vida, explorando palabras como subjetivo, social y familiar. Posteriormente, de 2009 a 2014, se introdujeron conceptos como emocional, sexualidad y cuidado de ancianos, lo que sugiere un interés creciente en aspectos más específicos de la salud y el bienestar. Un avance notable se observa entre 2015 y 2020 con la ampliación del enfoque para incluir palabras como violencia, maltrato y demencia, lo que refleja una comprensión más profunda de los desafíos que afectan la salud de las personas. Finalmente, en el período de 2021 a 2024, la aparición de palabras como adolescencia, cuidadores y derechos de los pacientes indica una mayor preocupación por la salud

mental y los derechos de las personas en diferentes etapas de la vida, lo que promete avances significativos en la promoción del bienestar integral. (Figura 4.6)

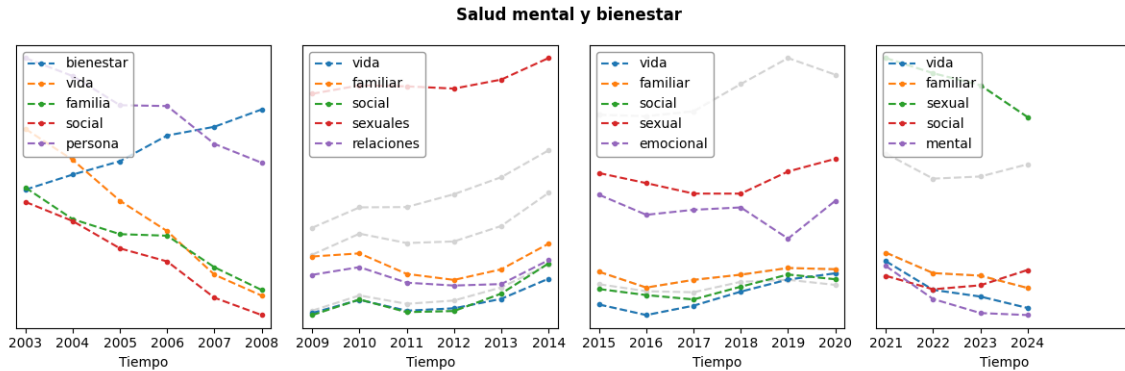


Figura 4.6: Gráficas de evolución de términos asociados al tópico 'Salud mental y bienestar' en distintos periodos.

### Enfermedades neurológicas y respiratorias (*topic#11*)

Inicialmente (2003-2008), se centraba en trastornos generales como epilepsia y asma, destacando síntomas cognitivos. Luego (2009-2014), se precisó el estudio de enfermedades neuromusculares y respiratorias, incluyendo aspectos genéticos y la relación con trastornos del sueño. Posteriormente (2015-2020), se expandió a enfermedades psiquiátricas y cognitivas, avanzando en diagnóstico y tratamiento. Finalmente (2021-2024), el enfoque se refinó hacia trastornos neurodegenerativos y enfermedades respiratorias obstructivas, utilizando protocolos médicos y publicando en plataformas científicas. En resumen, se pasó de trastornos generales a enfermedades específicas, abarcando aspectos genéticos, psiquiátricos y cognitivos, con avances en diagnóstico, tratamiento y difusión científica. (Figura 4.7)

### Casos clínicos y diagnósticos médicos (*topic#12*)

En los primeros años (2003-2008), se observa un énfasis en reportes de pacientes con afecciones abdominales, centrados en el diagnóstico clínico y quirúrgico. Se mencionan términos como manejo, biopsia, antecedentes, cavidad, diagnóstico simple, surgical. A partir de 2009, el enfoque se amplía hacia enfermedades raras y específicas, incorporando términos como pared, temporal, tumoral, arterial, lo que sugiere estudios más avanzados en patología abdominal. Entre 2015 y 2020, se introducen términos como quistes, intestino, hernia, ascitis, reflejando un interés en patologías

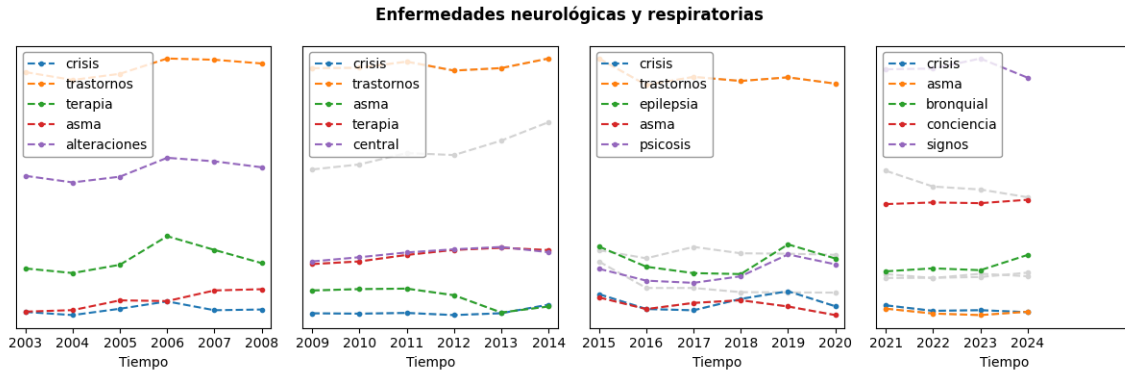


Figura 4.7: Gráficas de evolución de términos asociados al tópico 'Enfermedades neurológicas y respiratorias' en distintos periodos.

digestivas y peritoneales. Desde 2021 hasta 2024, la investigación se centra en condiciones más especializadas, como idiopathic venosa, lo que indica una profundización en patologías vasculares abdominales. (Figura 4.8)

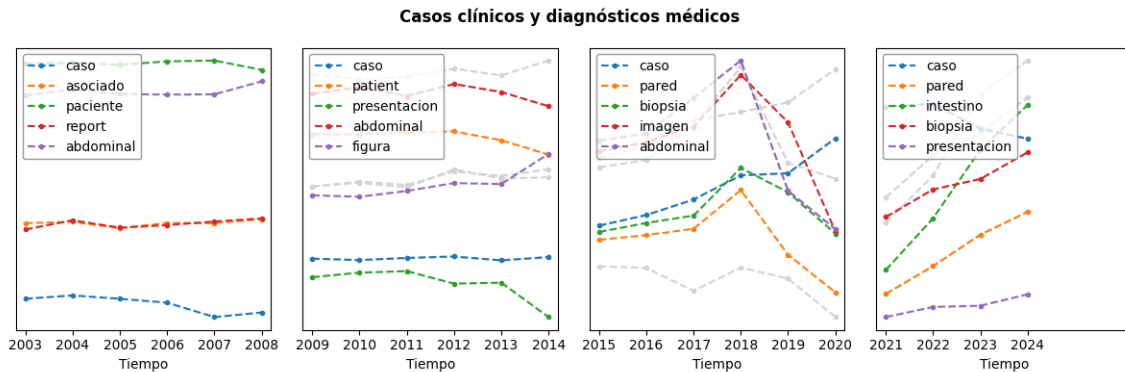


Figura 4.8: Gráficas de evolución de términos asociados al tópico 'Casos clínicos y diagnósticos médicos' en distintos periodos.

### Investigación en enfermedades renales e infecciosas (*topic#15*)

En los primeros años (2003-2008), la investigación se enfocaba en estudios con animales, particularmente en relación con el peso, la dosis y los efectos renales, utilizando ratas como modelo experimental. Luego, entre 2009 y 2014, el interés viró hacia las enfermedades infecciosas, haciendo hincapié en conceptos como infecciones, sepsis, plasma y sensibilidad. Más adelante, de 2015 a 2020, la problemática de la resistencia a los antibióticos cobró mayor relevancia, profundizando en el estudio



de organismos como el *Staphylococcus aureus* y la *Acinetobacter*. Finalmente, en el período 2021-2024, la investigación se expandió para abarcar enfermedades autoinmunes, incluyendo el estudio de anticuerpos, hepatitis y artritis reumatoide. (Figura 4.9)

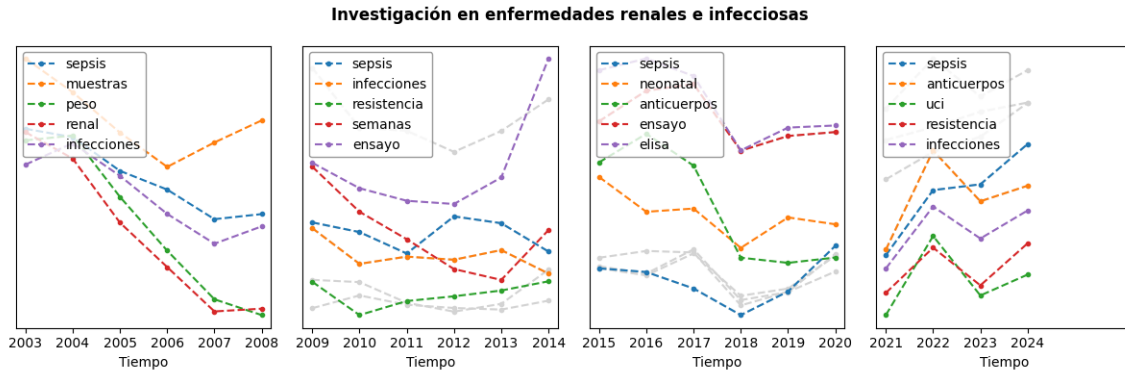


Figura 4.9: Gráficas de evolución de términos asociados al tópico 'Investigación en enfermedades renales e infecciosas' en distintos periodos.

### Factores de riesgo cardiovascular (*topic#19*)

Inicialmente (2003-2008), se centraba en factores de riesgo tradicionales como hipertensión, diabetes y obesidad. Luego, entre 2009 y 2014, el enfoque se amplió para incluir enfermedades cerebrovasculares, con términos como ictus, colesterol y aterosclerosis. Posteriormente, de 2015 a 2020, se exploró la conexión entre salud mental y enfermedades cardiovasculares, abarcando suicidio, trastornos mentales e intentos suicidas. Finalmente, desde 2021 hasta 2024, la investigación se ha extendido a la salud ósea y neonatal, incluyendo cadera, displasia y mortalidad prenatal. (Figura 4.10)

### Odontología y ortodoncia (*topic#20*)

En una primera etapa (2003-2008), la atención se centraba en la cirugía oral, con referencias a implantes dentales y estabilidad ósea. Posteriormente, entre 2009 y 2014, el foco se desplazó hacia los tratamientos correctivos, como la ortodoncia, el anclaje y aspectos oclusales. Luego, de 2015 a 2020, se observaron avances en diagnóstico y planificación, introduciendo conceptos como la predicción del crecimiento dental, la endodoncia y el hueso alveolar. Finalmente, en el período 2021-2024, la investigación se orientó hacia la incorporación de tecnología, enfatizando en fracturas mandibulares y el uso de modelos computacionales en la práctica odontológica. (Figura 4.11)

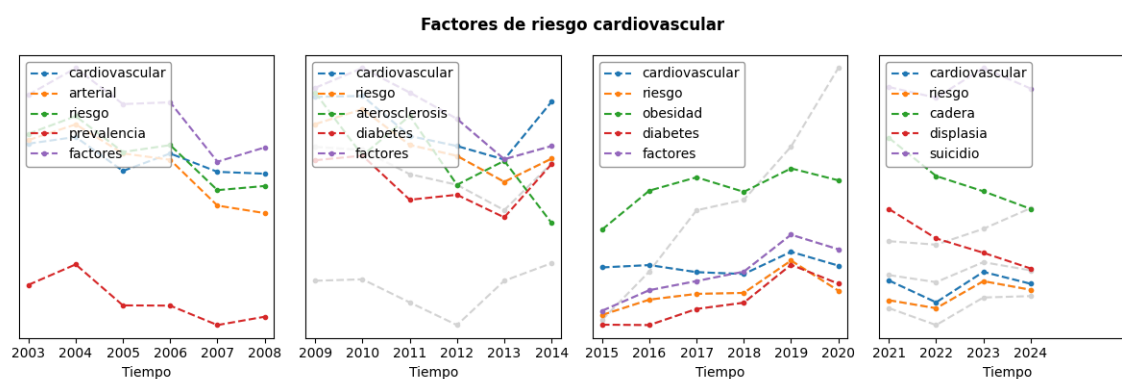


Figura 4.10: Gráficas de evolución de términos asociados al tópico 'Factores de riesgo cardiovascular' en distintos periodos.

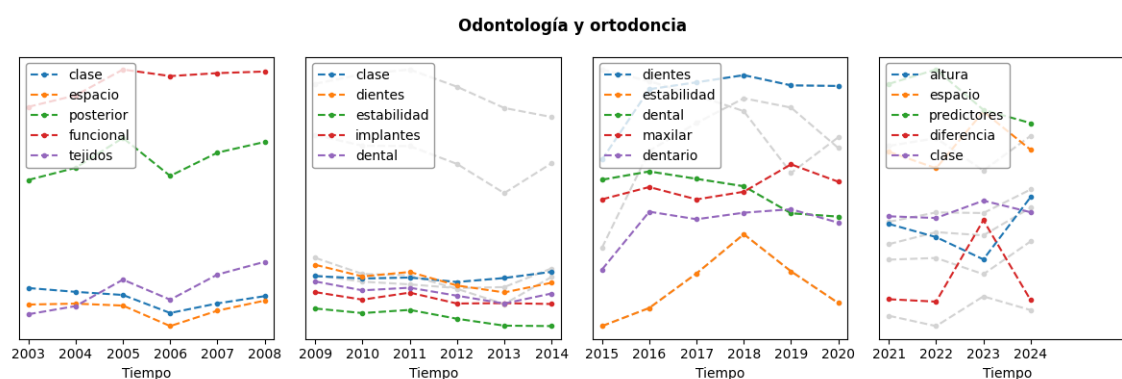


Figura 4.11: Gráficas de evolución de términos asociados al tópico 'Odontología y ortodoncia' en distintos periodos.

# Conclusiones

El presente estudio ha explorado y comparado el desempeño de los modelos de tópicos dinámicos DETM y CFDTM en el análisis de los artículos publicados en la Revista de Ciencias Médicas de La Habana. A lo largo del trabajo, se han realizado experimentos extensivos con distintas configuraciones de hiperparámetros con el objetivo de determinar cuál de los dos modelos logra representar de manera más efectiva la evolución de los temas en el tiempo.

Los resultados obtenidos indican que, si bien ambos modelos presentan un rendimiento similar en términos de coherencia de tópicos, CFDTM demuestra una ventaja significativa en términos de diversidad temática. La capacidad de este modelo para extraer tópicos más diferenciados y menos redundantes sugiere que representa mejor los cambios temáticos a lo largo del tiempo. Asimismo, los experimentos de clasificación y agrupamiento de documentos evidencian que CFDTM permite una mejor organización de los textos en función de sus temáticas, lo que facilita la identificación de tendencias emergentes en la investigación médica.

Otro aspecto relevante del estudio ha sido la búsqueda de configuraciones óptimas de hiperparámetros, lo que ha permitido mejorar notablemente el rendimiento de ambos modelos. La selección adecuada de estos parámetros ha resultado crucial para maximizar la calidad de los tópicos extraídos y garantizar la estabilidad del entrenamiento.

En conclusión, los hallazgos de esta investigación sugieren que el modelo CFDTM es más adecuado que DETM para analizar la evolución de los temas en revistas científicas. Su capacidad para generar tópicos más diversos y estructurados lo convierte en una herramienta valiosa para el estudio de la dinámica temática en publicaciones científicas especializadas.

# Recomendaciones

A partir de los resultados obtenidos en este estudio, se proponen varias líneas de trabajo futuro que podrían contribuir al desarrollo y la optimización de los modelos de tópicos dinámicos.

En primer lugar, sería interesante explorar la fusión del modelo CFDTM con TopicGPT (Pham et al. 2023) en lugar de LDA para la captura de tópicos en un momento específico en el tiempo. Dado que TopicGPT ha demostrado capacidades avanzadas en la generación de representaciones semánticas, su integración con CFDTM podría mejorar la calidad y precisión de los tópicos extraídos, proporcionando una mejor interpretación de la evolución temática en conjuntos de datos de naturaleza dinámica.

Asimismo, se recomienda la creación de una herramienta que, una vez entrenados los modelos de tópicos dinámicos, sea capaz de reconstruir visualmente la evolución de los tópicos a lo largo del tiempo. Este tipo de herramienta facilitaría la interpretación y el análisis de tendencias en investigaciones científicas, ofreciendo a los especialistas una forma más intuitiva de comprender la dinámica de los temas abordados en la literatura académica.

Finalmente, se sugiere ampliar la evaluación de los modelos incluyendo métricas adicionales que midan el impacto práctico de los tópicos extraídos en tareas específicas, como la recuperación de información y la recomendación de artículos científicos. La combinación de enfoques cualitativos y cuantitativos en la evaluación contribuiría a una mejor selección del modelo más adecuado según el propósito del análisis.

# Bibliografía

- Adji B. Dieng, D. M. B., Francisco J. R. Ruiz. (2019). The Dynamic Embedded Topic Model. <https://arxiv.org/abs/1907.05545> (vid. págs. 2, 6, 8).
- Song, D. J. C. Z. Y. (2023). *Probabilistic Topic Models. Foundation and Application*.
- Wu, X., Dong, X., Pan, L., Nguyen, T., & Luu, A. T. (2024). Modeling Dynamic Topics in Chain-Free Fashion by Evolution-Tracking Contrastive Learning and Unassociated Word Exclusion. <https://arxiv.org/abs/2405.17957> (vid. págs. 2, 8).
- Wu, X., Nguyen, T., & Luu, A. T. (2024). A Survey on Neural Topic Models: Methods, Applications, and Challenges. *ArXiv, abs/2401.15351*. <https://api.semanticscholar.org/CorpusID:267297321> (vid. pág. 7).
- Wu, X., Pan, F., & Luu, A. T. (2023). Towards the TopMost: A Topic Modeling System Toolkit. *ArXiv, abs/2309.06908*. <https://api.semanticscholar.org/CorpusID:261705592> (vid. págs. 10, 11).

# Referencias

- Adji B. Dieng, D. M. B., Francisco J. R. Ruiz. (2019). The Dynamic Embedded Topic Model. <https://arxiv.org/abs/1907.05545> (vid. págs. 2, 6, 8).
- Arthur P. Dempster, D. B. R., Nan M. Laird. (1977). Maximum likelihood from incomplete data via the EM - algorithm plus discussions on the paper. <https://api.semanticscholar.org/CorpusID:4193919> (vid. pág. 4).
- Baevski, A., Hsu, W.-N., Xu, Q., Babu, A., Gu, J., & Auli, M. (2022). data2vec: A General Framework for Self-supervised Learning in Speech, Vision and Language. *ArXiv, abs/2202.03555*. <https://api.semanticscholar.org/CorpusID:246652264> (vid. pág. 7).
- Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*, 112(518), 859-877. <https://doi.org/10.1080/01621459.2017.1285773> (vid. pág. 6).
- Blei, D. M., Ng, A., & Jordan, M. I. (2003). Latent Dirichlet Allocation. <https://api.semanticscholar.org/CorpusID:3177797> (vid. págs. 4, 6).
- Blei & Lafferty. (2006). Dynamic topic models. <https://doi.org/10.1145/1143844.1143859> (vid. págs. 6, 8).
- Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2016). Enriching Word Vectors with Subword Information. *arXiv preprint arXiv:1607.04606* (vid. pág. 10).
- Cvejoski, K., Sánchez, R. J., & Ojeda, C. (2023). Neural Dynamic Focused Topic Model. <https://arxiv.org/abs/2301.10988> (vid. pág. 7).
- Deerwester, S. C., Dumais, S. T., Landauer, T. K., Fumas, G. W., & Beck, L. L. (1988). Improving information retrieval using latent semantic indexing. <https://api.semanticscholar.org/CorpusID:59739393> (vid. pág. 4).
- Dieng, A. B., Ruiz, F. J. R., & Blei, D. M. (2019). Topic Modeling in Embedding Spaces. *Transactions of the Association for Computational Linguistics*, 8, 439-453. <https://api.semanticscholar.org/CorpusID:195886143> (vid. pág. 6).
- Gershman, S. J., & Goodman, N. D. (2014). Amortized Inference in Probabilistic Reasoning. *Cognitive Science*, 36. <https://api.semanticscholar.org/CorpusID:924780> (vid. pág. 6).

- Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9, 1735-1780. <https://api.semanticscholar.org/CorpusID:1915014> (vid. pág. 6).
- Hoffman, M., Blei, D. M., Wang, C., & Paisley, J. (2013). Stochastic Variational Inference. <https://arxiv.org/abs/1206.7051> (vid. pág. 6).
- Hofmann, T. (1999). Probabilistic Latent Semantic Analysis. *Conference on Uncertainty in Artificial Intelligence*, 289-296. <https://api.semanticscholar.org/CorpusID:653762> (vid. pág. 4).
- Jordan, M. I., Ghahramani, Z., Jaakkola, T., & Saul, L. K. (1999). An Introduction to Variational Methods for Graphical Models. *Machine Learning*, 37, 183-233. <https://api.semanticscholar.org/CorpusID:2073260> (vid. pág. 6).
- Kingma, D. P., & Welling, M. (2022). Auto-Encoding Variational Bayes. <https://arxiv.org/abs/1312.6114> (vid. pág. 8).
- Li, W., & McCallum, A. (2006). Pachinko allocation: DAG-structured mixture models of topic correlations. *Proceedings of the 23rd international conference on Machine learning*, 577-584. <https://api.semanticscholar.org/CorpusID:13160178> (vid. pág. 5).
- Lin, Z., Feng, M., dos Santos, C. N., Yu, M., Xiang, B., Zhou, B., & Bengio, Y. (2017). A Structured Self-attentive Sentence Embedding. <https://arxiv.org/abs/1703.03130> (vid. pág. 7).
- Miyamoto, N., Isonuma, M., Takase, S., Mori, J., & Sakata, I. (2023, julio). Dynamic Structured Neural Topic Model with Self-Attention Mechanism. En A. Rogers, J. Boyd-Graber & N. Okazaki (Eds.), *Findings of the Association for Computational Linguistics: ACL 2023* (pp. 5916-5930). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.findings-acl.366> (vid. págs. 7, 8).
- Murphy, K. P. (2012). Machine learning - a probabilistic perspective. *Adaptive computation and machine learning series*. <https://api.semanticscholar.org/CorpusID:17793133> (vid. pág. 5).
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., ... Chintala, S. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. *ArXiv, abs/1912.01703*. <https://api.semanticscholar.org/CorpusID:202786778> (vid. pág. 11).
- Perrone, V., Jenkins, P. A., Spano, D., & Teh, Y. W. (2016). Poisson Random Fields for Dynamic Feature Models. <https://arxiv.org/abs/1611.07460> (vid. pág. 7).
- Pham, C. M., Hoyle, A. M., Sun, S., & Iyyer, M. (2023). TopicGPT: A Prompt-based Topic Modeling Framework. *ArXiv, abs/2311.01449*. <https://api.semanticscholar.org/CorpusID:272144718> (vid. pág. 29).

- Rahimi, H., Naacke, H., Constantin, C., & Amann, B. (2023). ANTM: An Aligned Neural Topic Model for Exploring Evolving Topics. <https://arxiv.org/abs/2302.01501> (vid. pág. 7).
- Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the Space of Topic Coherence Measures. *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*. <https://api.semanticscholar.org/CorpusID:7743332> (vid. pág. 17).
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023). Attention Is All You Need. <https://arxiv.org/abs/1706.03762> (vid. pág. 7).
- Wang, X., & McCallum, A. (2006). Topics over time: a non-Markov continuous-time model of topical trends. *Knowledge Discovery and Data Mining*, 424-433. <https://api.semanticscholar.org/CorpusID:207160148> (vid. pág. 5).
- Williamson, S., Wang, C., Heller, K. A., & Blei, D. M. (2010). The IBP Compound Dirichlet Process and its Application to Focused Topic Modeling. *International Conference on Machine Learning*. <https://api.semanticscholar.org/CorpusID:6023511> (vid. pág. 7).
- Wu, X., Dong, X., Pan, L., Nguyen, T., & Luu, A. T. (2024). Modeling Dynamic Topics in Chain-Free Fashion by Evolution-Tracking Contrastive Learning and Unassociated Word Exclusion. <https://arxiv.org/abs/2405.17957> (vid. págs. 2, 8).
- Zhang, D. C., & Lauw, H. (2022, 17–23 Jul). Dynamic Topic Models for Temporal Document Networks. En K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu & S. Sabato (Eds.), *Proceedings of the 39th International Conference on Machine Learning* (pp. 26281-26292, Vol. 162). PMLR. <https://proceedings.mlr.press/v162/zhang22n.html> (vid. pág. 6).



# Anexos

## Lista completa de tópicos extraídos (2003 - 2024)

### Topic#1

2003: datos centros vigilancia red publicaciones oms casos ecuador ocurrencia instituciones  
2004: datos centros vigilancia red publicaciones oms casos ocurrencia ecuador instituciones  
2005: datos vigilancia centros casos oms red publicaciones ocurrencia ecuador impacto  
2006: datos vigilancia centros casos publicaciones oms red ocurrencia ecuador autores  
2007: datos vigilancia casos centros oms publicaciones red ocurrencia impacto ecuador  
2008: datos vigilancia casos centros oms red publicaciones ocurrencia impacto cantidad  
2009: datos vigilancia casos oms centros red publicaciones impacto ocurrencia ecuador  
2010: datos vigilancia centros casos oms red publicaciones impacto ocurrencia cantidad  
2011: datos vigilancia centros casos oms red publicaciones impacto cantidad instituciones  
2012: datos casos vigilancia centros oms red publicaciones impacto cantidad cuba  
2013: datos centros vigilancia casos oms red publicaciones impacto instituciones cuba  
2014: centros datos red vigilancia casos oms instituciones integrado impacto intoxicaciones  
2015: centros red datos casos vigilancia oms intoxicaciones integrado instituciones dorta  
2016: dorta contreras angiostrongylus citaciones centros cantonensis red casos datos publicado  
2017: dorta contreras datos centros casos red citaciones oms angiostrongylus cuba  
2018: datos casos centros oms red dorta contreras cuba publicaciones vigilancia  
2019: datos casos centros red oms dorta dengue impacto vigilancia contreras  
2020: casos medidas datos coronavirus covid nueva australia news oms national  
2021: datos pandemia data covid casos impacto infectados vacunas coronavirus susceptibles  
2022: covid competitividad san gobiernos regional ica mortality desigualdad almeida wwwrevhabanerasldcu  
2023: datos covid health digital modelos ruido vacuna hospitales pandemia research  
2024: datos health covid legal pandemia registros citado download vacuna research

## Topic#2

2003: municipio rev negativos mujeres revista adultos secundaria materna mayores ancianos  
2004: municipio rev negativos revista mujeres secundaria adultos materna tabaquismo medicamentos  
2005: municipio rev negativos revista mujeres tabaquismo secundaria vih medicamentos sida  
2006: municipio rev revista negativos mujeres secundaria vih medicamentos tabaquismo sida  
2007: municipio rev revista mujeres tabaquismo arttext negativos http script scielophp  
2008: municipio rev revista mujeres negativos tabaquismo

## Topic#3

2003: presentation descrito regiones cases variedad cuadro reportes caracteriza derecho reportado  
2004: presentation descrito regiones cases variedad cuadro reportes caracteriza derecho reportado  
2005: presentation regiones descrito cases variedad cuadro derecho reportes caracteriza cancer  
2006: presentation regiones descrito cases variedad cuadro reportes caracteriza cancer derecho  
2007: presentation regiones descrito cases variedad cancer reportes cuadro caracteriza derecho  
2008: presentation regiones descrito cases variedad reportes cancer caracteriza cuadro derecho  
2009: presentation regiones descrito variedad cases tumores reportes cuadro derecho caracteriza  
2010: presentation regiones descrito cases variedad tumores cuadro reportes derecho caracteriza  
2011: presentation regiones descrito variedad cases tumores reportes cuadro caracteriza two  
2012: presentation regiones descrito variedad cases tumores reportes cuadro two caracteriza  
2013: presentation descrito regiones variedad cases tumores cuadro reportes two tumor  
2014: presentation regiones descrito variedad tumores cases von tumor cuadro tos  
2015: presentation regiones descrito variedad tumores cases tumor von cuadro derecho  
2016: presentation regiones variedad descrito tumores cases von tumor when cuadro  
2017: presentation regiones variedad descrito tumores cases von cuadro tumor tos  
2018: presentation regiones variedad descrito cases von tumores cuadro aparecen tos  
2019: presentation regiones variedad descrito cases von tumores cuadro aparecen reportes  
2020: presentation regiones variedad descrito cases tumores von cuadro aparecen reportes  
2021: presentation regiones cases variedad descrito tumores aparecen cuadro von reportes  
2022: presentation regiones cases variedad descrito aparecen reportes when von cuadro  
2023: presentation regiones cases descrito variedad when reportes aparecen cuadro von  
2024: presentation regiones cases descrito variedad when aparecen reportes cuadro von

## Topic#4

2003: proceso educativo propuesta contenidos plan profesional habilidades objetivos profesores trabajo  
2004: proceso educativo profesional propuesta plan contenidos habilidades aprendizaje profesores trabajo

2005: proceso educativo preguntas propuesta estrategia contenidos expertos profesional objetivos plan  
 2006: proceso preguntas contenidos estrategia objetivos propuesta plan educativo profesional habilidades  
 2007: contenidos proceso plan propuesta estrategia preguntas objetivos profesional curricular educativo  
 2008: contenidos proceso plan propuesta estrategia preguntas objetivos curricular profesional carreras  
 2009: contenidos proceso plan propuesta profesional estrategia objetivos curricular habilidades profesionales  
 2010: contenidos proceso plan propuesta profesional curricular habilidades educativo aprendizaje trabajo  
 2011: contenidos plan proceso propuesta curricular habilidades profesional asignaturas disciplina educativo  
 2012: contenidos proceso plan propuesta profesional habilidades aprendizaje curricular educativo asignaturas  
 2013: contenidos proceso plan profesional propuesta habilidades trabajo educativo estudiantes aprendizaje  
 2014: contenidos proceso asignaturas plan estudiantes profesional propuesta aprendizaje estudiante profesores  
 2015: competencias aprendizaje proceso propuesta profesional docente estudiantes estudiante especialidad profesores  
 2016: competencias aprendizaje proceso docente propuesta profesional estudiante estudiantes especialidad profesores  
 2017: aprendizaje proceso docente competencias profesional propuesta habilidades educativo especialidad estudiante  
 2018: aprendizaje educativo docente proceso competencias profesional habilidades propuesta especialidad docentes  
 2019: aprendizaje educativo proceso docente competencias profesional habilidades propuesta especialidad estudiantes  
 2020: educativo aprendizaje proceso competencias profesional docente estudiantes docentes propuesta habilidades  
 2021: aprendizaje educativo proceso profesional competencias estudiantes docentes docente propuesta habilidades  
 2022: aprendizaje proceso educativo competencias profesional estudiantes docentes expertos propuesta contenido  
 2023: aprendizaje proceso profesional competencias educativo estudiantes competencia expertos docentes estudiante  
 2024: aprendizaje proceso profesional competencias educativo estudiantes competencia expertos especialistas docentes

## Topic#5

2003: operada frontales pujol livingstone testicular excluidas tanzania predisponente meningitidis lucha contra  
 2004: operada frontales livingstone tierras obrera filas cumbre pujol quiera tanzania  
 2005: operada reach impulso nombramiento accepted desea mid antiinflamatorio alerm claude  
 2006: objetivo operada acquired excelencia maleficencia programa desea habilidades alto glosario  
 2007: objetivo excelencia operada prof desee obrera propiciado amaro ujc arrojaron  
 2008: excelencia objetivo operada prof desea amaro obrera maleficencia darse  
 2009: excelencia objetivo operada prof desea gratuita excluidas fortaleza stand  
 2010: operada excluidas comparativa application dedican viejos organized ideales convocados  
 2011: excelencia operada gratuita editor delayed excluidas abdo odinofagia fortaleza promedio  
 2012: editor excelencia operada delayed odinofagia abdo domingo pioneros aprobados excluidas  
 2013: excelencia editor sanitarios exploratorio sustenta reviews ampliar office labores entrega  
 2014: excelencia editor sustenta sanitarios office desarrollada sede continuar labores talleres  
 2015: excelencia editor sustenta desarrollada sanitarios grupal office pupo proyectos desea  
 2016: excelencia desarrollada editor sanitarios office grupal sustenta programa pupo discusiones  
 2017: excelencia editor sanitarios desarrollada office lista pupo programa proyecciones proyectos  
 2018: excelencia editor desarrollada sanitarios lista office proyecciones pupo demandas ambos sexos

2019: editor excelencia sanitarios lista desarrollada proyecciones ambos sexos proyectos mayor prevalencia demandas  
 2020: editor lista sanitarios excelencia desarrollada registrado proyecciones office redes ambos sexos  
 2021: editor registrado sanitarios lista excelencia office medicamentos desarrollada confirmado ambos sexos  
 2022: editor lista sanitarios registrado excelencia redes desarrollada referente municipios exploratorio  
 2023: editor excelencia lista registrado sanitarios desarrollada redes municipios office demandas  
 2024: editor lista registrado excelencia sanitarios desarrollada redes comorbidities moderate institucional

## Topic#6

2003: relation agua area muestra reporta genes observan when probablemente afectadas  
 2004: relation agua area muestra genes reporta observan when probablemente principalmente  
 2005: relation agua area muestra observan reporta genes probablemente when nose  
 2006: relation agua area observan muestra reporta genes probablemente carcinoma nose  
 2007: relation agua area observan reporta muestra genes probablemente carcinoma utilizada  
 2008: relation agua area observan muestra reporta genes carcinoma probablemente utilizada  
 2009: relation agua area observan genes muestra carcinoma probablemente reporta piel  
 2010: relation agua area observan genes muestra carcinoma probablemente reporta piel  
 2011: relation agua area observan genes carcinoma muestra receptores probablemente reporta  
 2012: relation agua area genes observan carcinoma receptores muestra probablemente philadelphia  
 2013: relation agua area genes observan receptores carcinoma atrofia philadelphia respiratory  
 2014: agua relation area observan genes carcinoma receptores atrofia philadelphia manos  
 2015: agua relation area observan genes carcinoma atrofia oncology padecimiento receptores  
 2016: agua relation area observan genes carcinoma arroyo environmental padecimiento lisa  
 2017: agua relation area arroyo hill genes carcinoma observan raza doble  
 2018: agua relation area hill arroyo doble contiene raza diarios genes  
 2019: relation agua area hill ambulatoria raza diarios aparecer von doble  
 2020: relation agua area hill revisada ambulatoria involucrados hormonas aparecer santa  
 2021: relation area hill agua ambulatoria nose calle crecimiento hormonas revisada  
 2022: relation agua area hill ambulatoria casos crecimiento nose becker calle  
 2023: relation hill agua casos crecimiento area ambulatoria comouna nose skin  
 2024: relation agua area hill casos crecimiento nose comouna locual ambulatoria

## Topic#7

2003: edades descriptivo complicaciones lesiones predominio seguido presentaron hallazgos universo proceder 2004: edades descriptivo complicaciones lesiones predominio seguido presentaron hallazgos universo proceder  
 2005: edades descriptivo complicaciones lesiones predominio seguido presentaron hallazgos universo proceder  
 2006: edades descriptivo complicaciones lesiones predominio proceder hallazgos presentaron seguido consulta

2007: edades descriptivo complicaciones predominio lesiones presentaron seguido hallazgos proceder consulta  
 2008: edades descriptivo complicaciones predominio lesiones presentaron seguido hallazgos proceder consulta  
 2009: edades descriptivo complicaciones lesiones predominio presentaron seguido hallazgos proceder consulta  
 2010: edades descriptivo complicaciones lesiones predominio presentaron seguido hallazgos consulta estadio  
 2011: edades complicaciones descriptivo lesiones predominio hallazgos presentaron seguido estadio proceder  
 2012: edades descriptivo complicaciones lesiones predominio hallazgos seguido presentaron proceder mama  
 2013: edades complicaciones descriptivo fractures fracturas lesiones predominio mandibulares temporomandibular hallazgos  
 2014: complicaciones edades lesiones descriptivo predominio hallazgos fracturas fractures temporomandibular lesions  
 2015: lesiones edades complicaciones descriptivo predominio hallazgos lesions estadio seguido presentaron  
 2016: lesiones edades descriptivo complicaciones predominio hallazgos lesions seguido estadio presentaron  
 2017: lesiones edades descriptivo complicaciones predominio hallazgos seguido lesions estadio urgencias  
 2018: edades lesiones descriptivo predominio complicaciones hallazgos seguido urgencias accidentes universo  
 2019: edades lesiones descriptivo predominio complicaciones hallazgos lesions accidentes urgencias universo  
 2020: edades descriptivo lesiones predominio complicaciones lesions seguido hallazgos urgencias universo  
 2021: descriptivo edades lesiones predominio complicaciones lesions seguido universo surg urgencias  
 2022: descriptivo edades lesiones predominio complicaciones lesions seguido universo urgencias descriptive  
 2023: descriptivo edades lesiones complicaciones predominio lesions seguido proceder hallazgos universo  
 2024: descriptivo edades lesiones lesions accidentes acorde proceder acudieron ambulatoria predominio

## Topic#9

2003: bienestar vida familia life social persona individuo quality sociales primaria  
 2004: bienestar vida familia social life persona individuo sociales primaria salud  
 2005: bienestar vida familia social persona life individuo sociales salud familiar  
 2006: vida bienestar familia social persona familiar sociales salud individuo life  
 2007: vida familia bienestar persona social sociales familiar life individuo salud  
 2008: vida familia bienestar social persona sociales familiar life relaciones salud  
 2009: vida familia social persona bienestar familiar life sociales incontinencia padres  
 2010: vida familia social persona familiar bienestar sociales life relaciones padres  
 2011: vida familia social familiar persona sociales life bienestar sexual relaciones  
 2012: vida familia familiar social persona sociales sexual padres life relaciones  
 2013: vida familia familiar social persona padres sociales sexual life relaciones  
 2014: vida familia social familiar persona sociales padres relaciones life sexual  
 2015: vida familia social familiar sociales persona relaciones life padres emocional  
 2016: vida social familia familiar sociales persona mental emocional life relaciones  
 2017: vida familia social familiar sociales persona relaciones sexual emocional cuidado  
 2018: vida familiar familia social sociales persona sexual mental relaciones cuidado  
 2019: vida familiar familia social sexual sociales mental persona relaciones cuidado  
 2020: vida familiar familia social sociales sexual mental persona cuidado relaciones

2021: vida familiar familia sexual social sociales mental persona relaciones cuidado  
2022: vida familiar familia sexual mental social sociales persona emocional life  
2023: vida familiar familia sexual mental social persona sociales life emocional  
2024: vida familiar familia sexual social persona life mental sociales emocional

## Topic#11

2003: crisis trastornos asma puede inicio pueden drogas alteraciones central terapia  
2004: crisis trastornos asma puede inicio pueden alteraciones drogas central terapia  
2005: crisis trastornos herencia debilidad alteraciones tooth sleep locus inicio neurol  
2006: crisis sleep trastornos herencia debilidad inicio locus vigilia neurol pueden  
2007: crisis trastornos inicio terapia herencia pueden puede debilidad sleep alteraciones  
2008: crisis terapia trastornos inicio pueden puede alteraciones herencia neurol genes  
2009: crisis puede trastornos pueden inicio alteraciones terapia neurol asma central  
2010: crisis trastornos puede pueden alteraciones inicio asma terapia drogas central  
2011: crisis trastornos alteraciones puede pueden inicio asma central drogas terapia  
2012: crisis trastornos puede alteraciones asma pueden inicio central drogas terapia  
2013: crisis trastornos puede asma alteraciones pueden central drogas inicio trastorno  
2014: crisis trastornos central puede alteraciones pueden asma vestibular drogas syndrome  
2015: crisis trastornos puede central pueden alteraciones asma drogas syndrome paciente  
2016: crisis trastornos epilepsia epilepsy psicosis alteraciones pueden puede central asma  
2017: crisis trastornos epilepsia epilepsy pueden psicosis puede alteraciones asma epilepsias  
2018: crisis trastornos epilepsia epilepsy puede pueden asma alteraciones epilepsias drogas  
2019: crisis trastornos asma puede alteraciones pueden memoria drogas central inicio  
2020: crisis trastornos asma puede alteraciones central pueden memoria drogas inicio  
2021: crisis trastornos puede asma central alteraciones pueden inicio diversas conciencia  
2022: crisis trastornos puede asma pueden central inicio alteraciones conciencia signos  
2023: asma crisis bronquial bronquitis aguda puede obstructivo manual pueden bronchial  
2024: crisis asma puede aguda pueden trastornos central diversas manual signos

## Topic#12

2004: caso asociado presenta paciente report abdominal patient manejo figura presentation  
2005: caso asociado presenta paciente abdominal report patient presentation figura manejo  
2006: caso presenta asociado paciente patient abdominal report presentation figura simple  
2007: caso asociado presenta abdominal report patient paciente presentation figura diagnosis  
2009: caso asociado presenta abdominal report patient presentation paciente figura case  
2013: caso case asociado presenta patient temporal february report presentation abdominal

2015: caso case presenta presentation patient asociado abdominal report figura pared  
 2016: caso case presentation patient pared presenta abdominal imagen asociado biopsia  
 2019: caso case pared presentation abdominal presenta patient imagen biopsia tumoral  
 2021: caso case abdominal pared presentation imagen patient biopsia asociado intestino  
 2022: caso abdominal pared case presentation intestino biopsia patient intestinal imagen  
 2024: caso abdominal pared case presentation intestino intestinal patient asociado report

## Topic#15

2003: peso muestras efecto elevada renal dosis grupo respuesta ratas indicador  
 2004: muestras peso efecto elevada renal dosis grupo respuesta ratas semanas  
 2005: muestras peso efecto renal elevada dosis grupo respuesta semanas significativas  
 2006: elevada muestras peso renal efecto dosis indicador nacer grupo respuesta  
 2007: muestras elevada renal peso efecto semanas dosis respuesta sepsis infecciones  
 2008: muestras peso renal elevada efecto sepsis dosis grupo infecciones semanas  
 2009: muestras peso renal efecto elevada sepsis dosis infecciones semanas grupo  
 2010: muestras peso sepsis renal efecto elevada infecciones dosis resistencia semanas  
 2011: muestras sepsis renal peso efecto infecciones elevada dosis resistencia grupo  
 2012: muestras sepsis renal peso efecto infecciones elevada grupo dosis semanas  
 2013: muestras sepsis renal efecto peso infecciones anti semanas ensayo elevada  
 2014: sepsis muestras renal infecciones anti efecto peso semanas resistencia controles  
 2015: sepsis muestras anti renal infecciones semanas efecto ensayo peso resistencia  
 2016: semanas sepsis hepatitis renal neonatal anti muestras ensayo infecciones efecto  
 2017: sepsis semanas anti muestras renal hepatitis neonatal infecciones ensayo efecto  
 2018: anti sepsis neonatal muestras ensayo anticuerpos semanas ensayos infecciones elisa  
 2019: anti sepsis ensayo muestras anticuerpos neonatal ensayos infecciones elisa controles  
 2020: anti sepsis muestras ensayo anticuerpos neonatal infecciones controles ensayos semanas  
 2021: sepsis anti uci grupo infecciones intensivos anticuerpos muestras resistencia efecto  
 2022: grupo placebo efecto ratas vascular corporal peso frente liso cord  
 2023: ratas grupo efecto peso desinfectantes fetal control min corporal animales  
 2024: corneal ratas peso grupo visual defecto efecto hla agudeza aborto

## Topic#18

2004: raza operada frontales excluidas oportunista dedicadas metodo relatively debieron coagulation  
 2005: operada metodo asigificant raza relatively decided square excluidas problema investigative  
 2006: square bibliography operada adequate mesenchymal medications given practitioners proyectos odinofagia  
 2007: square mesenchymal operada proyectos medications bibliography odinofagia specialist excluidas focused

2008: operada square mesenchymal excluidas proyectos odinofagia relatively asignificant bibliography specialist  
 2009: operada proyectos garganta relatively excluidas odinofagia square mesenchymal financieros dedicadas  
 2011: operada garganta odinofagia excluidas frontales proyectos financieros relatively dedicadas square  
 2012: garganta odinofagia operada excluidas square frontales relatively proyectos batalla coordinadores  
 2013: editor sanitarios garganta square oficial teresa odinofagia coordinadores codigo proyectos  
 2014: sanitarios editor oficial square garganta odinofagia proyectos coordinadores teresa relatively  
 2015: editor sanitarios square garganta oficial estimado odinofagia relatively coordinadores proyectos  
 2016: estimado sanitarios editor square garganta oficial odinofagia teresa proyectos relatively  
 2017: estimado sanitarios square garganta odinofagia editor lista coordinadores publicar frontales  
 2018: sanitarios estimado garganta lista odinofagia square relatively frontales coordinadores dedicadas  
 2019: estimado sanitarios odinofagia square garganta proyectos relatively excluidas frontales coordinadores  
 2020: sanitarios estimado relatively odinofagia proyectos lista frontales square operada dedicadas  
 2021: square frontales excluidas prevention relatively operada morbi odinofagia sanitarios pujol  
 2022: sanitarios square morbi odinofagia relatively garganta lista excluidas frontales estimado  
 2023: square relatively estimado prevention dedicadas frontales garganta sanitarios operada  
 2024: estimado garganta lista odinofagia relatively square sanitarios proyectos dedicadas cubanamed

## Topic#19

2003: arterial cardiovascular riesgo risk muerte prevalencia factores factors diabetes cifras  
 2004: arterial cardiovascular riesgo risk prevalencia muerte factores factors diabetes causas  
 2005: arterial cardiovascular riesgo muerte prevalencia risk factores factors diabetes cifras  
 2006: arterial cardiovascular riesgo muerte prevalencia risk factores factors causas cifras  
 2007: arterial cardiovascular riesgo muerte risk cmd retrieve queryfcgi list prevalencia  
 2008: arterial cardiovascular riesgo muerte risk cmd prevalencia factores queryfcgi factors  
 2009: arterial cardiovascular risk riesgo prevalencia factors factores cerebrovascular stroke muerte  
 2010: arterial cardiovascular risk riesgo cifras factors prevalencia aterosclerosis factores diabetes  
 2011: arterial cardiovascular risk riesgo cifras factors aterosclerosis prevalencia factores circunferencia  
 2013: arterial cardiovascular riesgo risk cifras factores factors prevalencia diabetes aterosclerosis  
 2015: arterial cardiovascular riesgo risk factores cifras factors prevalencia aterosclerosis obesidad  
 2016: riesgo arterial cardiovascular risk factores factors prevalencia cifras diabetes muerte  
 2021: cardiovascular riesgo prevalencia risk factores arterial suicidio muerte diabetes causas  
 2022: cardiovascular riesgo prevalencia arterial muerte factores risk causas enfermedades suicidio  
 2023: cadera cardiovascular prevalencia causas riesgo displasia arterial muerte discapacidad risk  
 2024: cardiovascular prevalencia arterial riesgo cadera muerte risk causas factores diabetes



## Topic#20

2003: espacio clase superiores posterior inferior funcional promedio diferencia tratamiento tejidos  
2004: clase espacio superiores posterior inferior funcional promedio diferencia tratamiento anterior  
2005: clase espacio superiores posterior inferior funcional promedio tratamiento tejidos treatment  
2006: clase espacio posterior superiores tejidos facial dientes funcional anclaje aparatos  
2007: clase espacio superiores posterior funcional inferior tejidos dientes facial estabilidad  
2008: clase espacio superiores inferior posterior funcional promedio tejidos dientes facial  
2009: clase espacio dientes estabilidad superiores inmediata funcional posterior implante implantes  
2010: clase espacio dientes estabilidad superiores inmediata implante implantes inferior posterior  
2011: clase espacio dientes superiores estabilidad funcional inferior par inmediata promedio  
2012: clase dientes estabilidad espacio superiores dental implantes dentales inmediata implante  
2013: clase dientes estabilidad diente dental superiores dent espacio implantes maxilar  
2014: clase dientes estabilidad diente superiores facial dental oclusal dent molar  
2015: estabilidad dientes diente clase dental dent hueso dentario maxilar dentaria  
2017: estabilidad dientes dental diente dent maxilar hueso alveolar clase dentario  
2018: estabilidad dental dientes diente alveolar maxilar dent hueso dentaria dentario  
2019: estabilidad dental dientes diente alveolar maxilar hueso dent dentaria altura  
2020: estabilidad dientes dental diente alveolar maxilar hueso dent dentaria dentario  
2021: altura significativa inferior positiva distancia izquierdo rey crecimiento pie extremo  
2022: altura distancia inferior espacio rey significativa extremo izquierdo positiva rama  
2023: johnston espacio reales ramos diferencia superiores inferiores utilizando predictores significativa  
2024: johnston espacio reales ramos superiores inferiores predictores utilizando diferencia clase