Universidad de La Habana
Facultad de Matemática y Computación

Título de la tesis

Autor:
Naomi Lahera Champagne

Tutores:
Dr. Luciano García Garrido

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

Fecha

github.com/username/repo

Dedicación

Agradecimientos

Agradecimientos

Opinión del tutor

Opiniones de los tutores

Resumen

Resumen en español

Abstract

Resumen en inglés

Índice general

Índice de figuras

Introducción

Los modelos 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 de este corpus, se identifican todas las palabras únicas, formando el vocabulario. El objetivo es descubrir los temas subyacentes en estos documentos. Cada tema se define por un conjunto de palabras del vocabulario y, a su vez, cada documento se puede asociar a una combinación de estos temas.

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 *Probabilistic Latent Semantic Analysis* (PLSA) (Hofmann 1999), propuesto por Thomas Hofmann, 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. En la inicialización, cada palabra se asigna independientemente a un tópico. 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.

LDA sentó las bases para modelos que exploran tanto la estructura jerárquica como la evolución temporal de los temas. Modelos como *Raphil* (Murphy 2012) y *Pachinko Allocation Model* (Li y McCallum 2006) representan temas en forma de árbol, mientras que *Topic Over Time* (Wang y McCallum 2006) captura cómo estos temas 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 tópicos tradicionales impulsaron la investigación hacia enfoques más sofisticados. Los modelos 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 temas latentes en un corpus de documentos de manera más eficiente y precisa.

Un ejemplo destacado de estos modelos son las redes neuronales tópicas basadas en Variational Autoencoders (VAEs). Los VAEs, en este contexto, constan de dos componentes principales: un codificador y un decodificador. El codificador se encarga de inferir la distribución de temas a partir de un documento dado, es decir, trata de determinar qué temas son los más relevantes para ese documento en particular. Por otro lado, el decodificador tiene la función inversa: a partir de una distribución de temas, genera un nuevo documento que refleje esos temas. La principal ventaja de los modelos tópicos neuronales basados en VAEs es su capacidad para manejar grandes conjuntos de datos y capturar relaciones complejas entre palabras y temas. Además, son altamente personalizables y pueden adaptarse a diferentes tipos de datos y tareas.

El campo de los modelos 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 tópicos eficientes en Cuba, no es posible comprender cómo evolucionan los temas de investigación en revistas científicas como la Revista de Ciencias Médicas de La Habana. Esta limitación dificulta la identificación tendencias y la toma de decisiones informadas sobre la investigación.

El empleo de modelos de tópicos dinámicos permite rastrear el comportamiento de tópicos latentes en una colección de documentos en los que influye el momento en el que estos fueron publicados. A diferencia de los enfoques estáticos, estos modelos rastrean cómo los temas surgen, se modifican o se desvanecen, modelando la evolución de su contenido. De esta manera, se identifican tendencias de investigación y se

comprende mejor la trayectoria del conocimiento en la publicación.

El objetivo principal de esta investigación es determinar cuál de los modelos tópicos dinámicos, *Dynamic Embedded Topic Model*(DETM) Wu, Dong et al. 2024 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 ca- lidad y diversidad de los temas identificados; posteriormente, proponer un modelo adecuado para el análisis dinámico de temas 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 tres capítulos principales: Estado del Arte, Implementación y Resultados. El primer capítulo presenta una revisión exhaustiva de los trabajos previos relacionados, mientras que los capítulos subsiguientes detallan el proceso de implementación y los hallazgos obtenidos.

Capítulo 1

Estado del Arte

Los Modelos 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 [1] 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)* (Adji B. Dieng 2019) que sigue un enfoque diferente representando a palabras y tópicos como *embeddings* [2]. 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 misma 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 de 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

[1] 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.

[2] Un *embedding* es el mapeo de una variable categórica a un vector numérico.

y son aparoximados 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, 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).

En 2022 fue publicado el artículo Zhang y Lauw 2022, 17–23 Jul donde se describen dos modelos 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, es posible afirmar 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* (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 ventana deslizante superpuesta para la agrupación temporal de documentos. Este algoritmo, 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, 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 cómo 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) (Cvejoski 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 desacoplamineto 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 neural (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 [3] 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. 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, donde para rastrear la evolución del tema con diferentes intensidades construye relaciones positivas y más importante aún, construye relaciones negativas que alientan a los tópicos dentro de un mismo espacio de tiempo a ser distintos, manteniendo la diversidad de tópicos y

[3] Los tópicos dentro de un segmento de tiempo se consediran repetitivos al presentar una semántica similar.

así aliviando el problema de los tópicos repetitivos. Por otro lado UWE encuentra las palabras más relacionadas de los tópicos en cada segmento de tiempo e identifica cuáles de ellas no pertenecen a esta 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 [4] con mejoras significantes repecto al resto de los modelos existentes.

En conclusión, los Modelos Tópicos Neuronales Dinámicos han evolucionado considerablemente, incorporando técnicas avanzadas como embeddings, redes neuronales y mecanismos de autoatención para superar las limitaciones de los enfoques tradicionales. Las innovaciones recientes, como el uso de estructuras en red, modelos alineados, mecanismos de desacoplamiento y enfoques sin cadenas de Makov, 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.

[4] TC: Topic Coherence. Métrica usada para medir la coherencia entre las palabras asignadas a un tópico por un modelo tópico.

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), 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 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, que consiste en el conjunto único 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

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 específicos: 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 particular.

Para la implementación de los modelos de tópicos dinámicos DETM y CFDTM, se utilizó el módulo de Python TopMost [1] (Wu, Pan y Luu 2023) que usa PyTorch (Paszke et al. 2019) como marco de trabajo de redes neuronales para modelos 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 adquitectura de TopMost se muestra en la Figura 2.1.

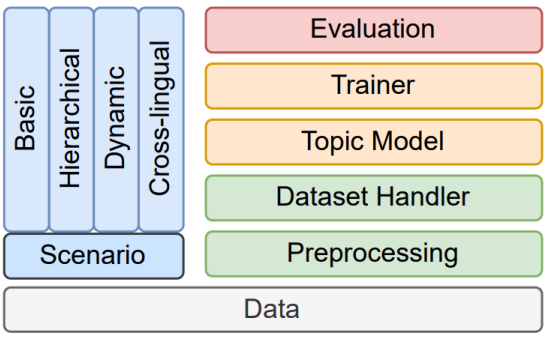

Figura 2.1: Arquitectura general de TOPMOST. Cubre los escenarios más comunes de modelado de temas y desacopla la carga de datos, la construcción de modelos, el entrenamiento de modelos y las evaluaciones en los ciclos de vida del modelado de tema. (Wu, Pan y Luu 2023)

El entrenamiento de los modelos de tópicos dinámicos DETM y CFDTM se llevó a

[1] https://github.com/bobxwu/topmost

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, GPU, lo que permite acelerar el proceso de entrenamiento de modelos complejos. Cada entorno de ejecución en Kaggle dispone de hasta 16 GB de RAM en el caso de las GPU y de hasta 30 GB de RAM en el caso de las CPU.

2.2.1. Modelo *Dynamic Embedding Topic Model (D-ETM)* (Adji B. Dieng 2019)

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.

Hiperparámetros

- **Número de tópicos:** Referente al número de tópicos latentes que el modelo intentará descubrir en el corpus. Un valor muy bajo puede mezclar temas distintos mientras que un valor muy alto puede granular un mismo tópico.
- **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.
- **Varianza de la Distribución Gaussiana para la Evolución de los Tópicos:** Controla la varianza del *random walk* en el espacio de *embeddings* de los tópicos.

2.2.2. Modelo *Chain-Free Dynamic Topic Model (CFDTM) (Wu, Dong et al. 2024)*

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

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

Hiperparámetros

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

- **Tasa de Aprendizaje:** Controla el tamaño de los pasos que da el optimizador durante el entrenamiento. Una tasa de aprendizaje muy alta puede causar inestabilidad en el entrenamiento, mientras que una muy baja puede llevar a una convergencia lenta.
- **Temperatura de *InfoNCE*:** Este parámetro controla la suavidad de la distribución de similitud en la función de pérdida *InfoNCE*. Afecta la forma en que el modelo distingue entre muestras positivas y negativas. Una temperatura baja hace que el modelo se centre más en las diferencias entre muestras, mientras que una temperatura alta suaviza las distinciones.
- **Umbral de Exclusión de Palabras No Asociadas:** Define el límite de la puntuación de asociación por debajo del cual una palabra se considera no asociada con un tópico y se excluye del cálculo de la pérdida. Un valor muy alto puede excluir demasiadas palabras, mientras que uno muy bajo puede no ser efectivo para eliminar el ruido.

Capítulo 3

Resultados

Resultados

Conclusiones

Conclusiones

Recomendaciones

Recomendaciones

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. 4, 7, 10).

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. 3, 6, 10).

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. 6).

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. 8, 9).

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. 4, 7, 10).

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. 1).

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. 5).

Blei, D. M., Ng, A., & Jordan, M. I. (2003). Latent Dirichlet Allocation. https://api.semanticscholar.org/CorpusID:3177797 (vid. págs. 1, 4).

Blei & Lafferty. (2006). Dynamic topic models. https://doi.org/10.1145/1143844.1143859 (vid. págs. 4, 7).

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2016). Enriching Word Vectors with Subword Information. *arXiv preprint arXiv:1607.04606* (vid. pág. 8).

Cvejoski, K., Sánchez, R. J., & Ojeda, C. (2023). Neural Dynamic Focused Topic Model. https://arxiv.org/abs/2301.10988 (vid. pág. 6).

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. 4).

Gershman, S. J., & Goodman, N. D. (2014). Amortized Inference in Probabilistic Reasoning. *Cognitive Science*, *36*. https://api.semanticscholar.org/CorpusID:924780 (vid. pág. 5).

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*, 1735-1780. https://api.semanticscholar.org/CorpusID:1915014 (vid. pág. 5).

Hoffman, M., Blei, D. M., Wang, C., & Paisley, J. (2013). Stochastic Variational Inference. https://arxiv.org/abs/1206.7051 (vid. pág. 5).

Hofmann, T. (1999). Probabilistic Latent Semantic Analysis. *Conference on Uncertainty in Artificial Intelligence*, 289-296. https://api.semanticscholar.org/CorpusID:653762 (vid. pág. 1).

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. 5).

Kingma, D. P., & Welling, M. (2022). Auto-Encoding Variational Bayes. https://arxiv.org/abs/1312.6114 (vid. pág. 6).

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. 2).

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. 5).

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. 5, 7).

Murphy, K. P. (2012). Machine learning - a probabilistic perspective. *Adaptive computation and machine learning series*. https://api.semanticscholar.org/CorpusID:17793133 (vid. pág. 2).

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. 9).

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. 6).

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. 5).

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. 5).

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. 2).

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. 6).

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. 3, 6, 10).

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. 5).