Naomi Lahera Champagne
Modelos Tópicos

Tarea No.3
Universidad de La Habana

Introducción

Los modelos tópicos son herramientas matemáticas y 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.

En el contexto de los modelos tópicos, el corpus se refiere al conjunto de documentos, mientras que el vocabulario denota todas las palabras únicas presentes en algún texto del corpus. El objetivo principal de un modelo tópico es identificar los tópicos latentes en la colección y asignar a cada elemento del corpus una proporción de estos tópicos. Cada tópico está compuesto por palabras del vocabulario del corpus.

Los enfoques clásicos incluyen modelos basados en factorización de matrices no negativas, como el *Latent Semantic Analysis* (LSA) [4] , que inicialmente define una matriz término-documento para representar el corpus y luego aplica descomposición en valores singulares (SVD). Mediante la descomposición $X = U\Sigma V^T$, U captura la relevancia de las palabras en los tópicos, V^T representa los documentos en el espacio temático y Σ es una matriz diagonal con valores singulares. Al truncar en función de un K (hiperparámetro) los valores singulares pequeños en Σ, junto con las columnas y filas correspondientes en U y V, 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 *Probabilistic Latent Semantic Analysis* (PLSA) [5], 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) [2] para optimizar los parámetros.

El modelo *Latent Dirichlet Allocation* (LDA) [3], 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 inspiró modelos especializados para escenarios no cubiertos por este enfoque. Ejemplos incluyen modelos jerárquicos, como *Raphil* [7] y *Pachinko Allocation Model* [6], que representan relaciones tópicas en varias capas mostrando los tópicos en forma arborea donde la raíz es el tópico más general, un mayor nivel en el árbol implica mayor especificidad del tópico llegando a ser las hojas conjuntos formados por una única palabra del corpus. Además están los modelos dinámicos como *Topic Over Time* [8], que capturan la evolución de los tópicos en documentos secuenciales, útiles en análisis de blogs, tweets y revistas.

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 tópicos con un grado considerable de eficiencia, demostrando su utilidad en diversos contextos. No obstante, aún persisten oportunidades para ampliar su aplicación a escenarios más diversos y específicos, como el análisis dinámico de temas en publicaciones científicas,

el seguimiento de tendencias en redes sociales o la exploración de datos en áreas especializadas. Profundizar en estos aspectos podría no solo potenciar el impacto de estos modelos, sino también contribuir al avance científico en contextos locales y globales.

En el contexto cubano, las revistas científicas desempeñan un papel crucial en la difusión del conocimiento y la construcción de comunidades académicas. Sin embargo, las herramientas actuales de análisis de texto en Cuba son estáticas y no permiten estudiar la evolución de los tópicos en el tiempo. Este análisis es clave para identificar tendencias, áreas emergentes y prioridades en la investigación nacional.

La presente investigación propone comparar el desempeño de modelos tópicos dinámicos *DETM* [1] y *CFDTM* [9] suministrándoles como dataset los artículos de la revista cubana *Ciencias Médicas*. Esto permitirá explorar cómo los modelos dinámicos pueden capturar la evolución de tópicos en un dominio especializado, destacando patrones relevantes y realizando ajustes específicos en los hiperparámetros para maximizar su eficacia. Además del ajuste de parámetros, se explorarán nuevas metodologías para enriquecer el modelo, como la incorporación de metadatos adicionales. Estos metadatos pueden incluir detalles como categorías temáticas específicas, áreas de investigación predominantes en cada artículo, o información temporal detallada. Agregar esta información puede fortalecer la capacidad del modelo para capturar de manera más precisa la evolución de los temas, ya que tendrá un mayor contexto sobre cada artículo y sus relaciones con otros dentro de la revista..

Este enfoque no solo busca validar la adaptabilidad de los modelos dinámicos a contextos especializados, sino también establecer un precedente para futuras aplicaciones en dominios únicos. Al final, se pretende desarrollar herramientas más versátiles para el análisis académico, capaces de representar con mayor precisión las características de los datos específicos y contribuir al avance del conocimiento en diversas áreas.

Esta tesis se desarrolla en la Facultad de Matemática y Computación de la Universidad de La Habana.

Referencias

[1] David M. Blei Adji B. Dieng Francisco J. R. Ruiz. *The Dynamic Embedded Topic Model*. 2019. arXiv: 1907.05545 [cs.CL]. URL: https://arxiv.org/abs/1907.05545.

[2] Donald B. Rubin Arthur P. Dempster Nan M. Laird. "Maximum likelihood from incomplete data via the EM - algorithm plus discussions on the paper". In: 1977. URL: https://api.semanticscholar.org/CorpusID:4193919.

[3] David M. Blei, A. Ng, and Michael I. Jordan. "Latent Dirichlet Allocation". In: 2003. URL: https://api.semanticscholar.org/CorpusID:3177797.

[4] Scott C. Deerwester et al. "Improving information retrieval using latent semantic indexing". In: 1988. URL: https://api.semanticscholar.org/CorpusID:59739393.

[5] Thomas Hofmann. "Probabilistic Latent Semantic Analysis". In: *Conference on Uncertainty in Artificial Intelligence*. 1999, 289–296. URL: https://api.semanticscholar.org/CorpusID:653762.

[6] Wei Li and Andrew McCallum. "Pachinko allocation: DAG-structured mixture models of topic correlations". In: *Proceedings of the 23rd international conference on Machine learning* (2006), 577–584. URL: https://api.semanticscholar.org/CorpusID:13160178.

[7] Kevin P. Murphy. "Machine learning - a probabilistic perspective". In: *Adaptive computation and machine learning series*. 2012. URL: https://api.semanticscholar.org/CorpusID:17793133.

[8] Xuerui Wang and Andrew McCallum. "Topics over time: a non-Markov continuous-time model of topical trends". In: *Knowledge Discovery and Data Mining*. 2006, pp. 424–433. URL: https://api.semanticscholar.org/CorpusID:207160148.

[9] Xiaobao Wu et al. *Modeling Dynamic Topics in Chain-Free Fashion by Evolution-Tracking Contrastive Learning and Unassociated Word Exclusion*. 2024. arXiv: 2405.17957 [cs.CL]. URL: https://arxiv.org/abs/2405.17957.

Bibliografía

[1] David M. Blei Adji B. Dieng Francisco J. R. Ruiz. *The Dynamic Embedded Topic Model*. 2019. arXiv: 1907.05545 [cs.CL]. URL: https://arxiv.org/abs/1907.05545.

[9] Xiaobao Wu et al. *Modeling Dynamic Topics in Chain-Free Fashion by Evolution-Tracking Contrastive Learning and Unassociated Word Exclusion*. 2024. arXiv: 2405.17957 [cs.CL]. URL: https://arxiv.org/abs/2405.17957.