Naomi Lahera Champagne
Modelos Tópicos

Tarea No.3
Universidad de La Habana

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.

Los métodos tradicionales, como el *Latente Semantic Analisis* (LSA) [3], emplean la factorización de matrices no negativas para revelar la estructura temática latente de un corpus. Inicialmente, se construye una matriz donde las filas representan términos y las columnas documentos. Luego, la *Single Value Descomposition* (SVD) descompone esta matriz en tres: U, una matriz de términos por tópicos; V^T, una matriz de documentos por tópicos; y Σ, una matriz diagonal que contiene los valores singulares. Al seleccionar los (K) valores singulares más grandes y sus correspondientes vectores singulares, se obtiene una aproximación de bajo rango de la matriz original. Los elementos de U y V^T indican, respectivamente, la relevancia de las palabras en los tópicos y la distribución de tópicos en los documentos.

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) [4], 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) [1] para optimizar los parámetros.

El modelo *Latent Dirichlet Allocation* (LDA) [2], 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* [6] y *Pachinko Allocation Model* [5] representan temas en forma de árbol, mientras que *Topic Over Time* [7] 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. Sin embargo, la ausencia de herramientas analíticas dinámicas limita la capacidad de los profesionales para comprender la evolución de los temas de investigación en revistas científicas como la revista de Ciencias Médicas de La Habana. Este trabajo de diploma busca abordar esta problemática aplicando modelos dinámicos de tópicos (DETM y CFDTM) para identificar y rastrear la evolución de los temas a lo largo del tiempo

con la premisa de que la incorporación de metadatos a estos modelos permitirá una representación más precisa y detallada de la evolución temática, revelando patrones subyacentes y facilitando la identificación de tendencias emergentes. Especificamente se pretende hacer una comparación en función del desempeño de ambos modelos con diferentes configuraciones de parámetros y evaluar su efectividad utilizando métricas como *Topic Coherence* y *Topic Diversity*. Se espera que los resultados de esta investigación contribuyan a desarrollar herramientas más sofisticadas para el análisis académico y a fortalecer la capacidad de los investigadores cubanos para identificar áreas de investigación prioritarias y anticipar futuras tendencias.

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.

Referencias

[1] 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.

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

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

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

[5] 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.

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

[7] 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.

Bibliografía