Naomi Lahera Champagne
Modelos Tòpicos

Tarea No.3
Universidad de La Habana

Introducción

Los modelos tópicos son modelos matemáticos o computacionales que permiten extraer la estructura temática latente en una colección de documentos. Siendo el corpus una colección de documentos llamaremos vocabulario al conjunto formado por cada palabra presente en alguno de los texto del corpus. Dada un corpus, un modelo tópico tiene el objetivo de extraer los tópicos latentes presentes en la colección y asignar a cada elementos del documento la proporción de cada tópico encontrado, donde cada tópico está compuesto por palabras del corpus.

La identificación de los tópicos latentes en una colección de documentos permite automatizar tareas como el análisis semántico, la recomendación de contenido. y la detección de sinonimia y polisemia.

Los enfoques convencionales de modelos tópicos incluyen modelos basados en factorización de matrices no negativas. Estos modelos descubren directamente los temas descomponiendo una matriz término-documento en dos matrices factoriales de bajo rango: una representa palabras y la otra documentos. El modelo Latente Semantic Analisis (LSA) (poner referencia) sigue este enfoque aplicando la descomposición en valores singulares (SVD) (poner referencia) a la matriz de ocurrencia palabra-documento (X) donde $a_{ij} = 1$ si la palabra número i del vocabulario aparece en el documento número j del corpus y $a_{ij} = 0$ en otro caso. Aplicando la descomposición SVD tenemos:$X = U\Sigma V^T$, donde U revela la importancia de las palabras en cada tema, mientras que V^T produce las representaciones de los documentos en el espacio temático, siendo Σ una matriz diagonal y los valores de su diagonal los valores singulares. En función del parámetro referente al número de tópicos (k) LSA trunca los valores singulares pequeños en Σ y sus filas y columnas correspondientes en U y V, para producir una representación con dimensiones reducidas. De esta forma de obtienen las proporciones de tópicos por documento y palabras por tópicos.

LSA fue comunmete usado en el campo de la recuperación de información. A pesar de esto la viabilidad de este modelo disminuye por necesidad de selección del parámetro K, el consumo de recursos computacionales y falta de expresividad de las represntaciones de los resultados ya que las matrices resultantes pueden tener valores positivos y negativo siendo contraintuitiva la interpretación de los valores. Para atacar estas debilidades de LSA surgió Probabilistic Latente Semantic Analisis (PLSA) (poner referencia), propuesto por Thomas Hofmann, un modelo tópico porbabilista que asume que cada documento del corpus es una distribución multinomial de tópicos latentes, siendo las

palabras, documentos y distribución de tópicos por documentos las variables de este modelo. PLSA se basa en la representación de las relaciones entre las variables como una Red Bayesiana, donde los tópicos son variables latentes y el resto son variables observadas (ver fig red bayesiana) y utiliza el algoritmo Expectation Maximization (EM) (poner referencia) para computar los resultados.

El enfoque probabilístico o bayesiano es también uno de los convencionales. Uno de los modelos clásicos más conocidos y eficientes es el Latent Dirichlet Allocation (LDA) propuesto por David M. Blei en el año 2003. Para el funcionamiento de este modelo se define formalmente un tópico como una distribución sobre un vocabulario predefunido y tiene en cuenta que en un documento pueden ocurrir varios tópicos de forma simultánea, teniendo como una de sus características principales que todos los documentso compartan los mismo K tópicos con prorporciones diferentes.

El proceso de asignación de tópicos en modelos probabilísticos se compone de tres fases: inicialización, reasignación y optimización. En la inicialización, cada palabra de un documento se asocia independientemente con un tópico, generando una distribución inicial donde los tópicos contienen conjuntos de palabras, permitiendo que una palabra esté en más de un tópico simultáneamente.

En la reasignación, se recalculan los tópicos de las palabras en los documentos considerando la frecuencia de uso de cada tópico en el documento y la frecuencia con que los tópicos contienen dichas palabras. Estos cálculos emplean recuentos iniciales y distribuciones multinomiales generadas por Dirichlet. Cada palabra se reasigna al tópico con mayor probabilidad condicional de pertenencia.

En la optimización, las reasignaciones se refinan iterativamente hasta alcanzar la convergencia, definida como un mínimo cambio en las asignaciones o un número fijo de iteraciones. El modelo utiliza probabilidades condicionales basadas en redes bayesianas para obtener una representación final optimizada de los tópicos, ajustada a la estructura de los documentos y palabras de la colección.

La presición de LDA propició que los investigadores desarrollaran modelos basándose en las asumpsiones y en el funcionamiento de LDA. Algunos de estos modelos fueron disennados ára ser aplicados en escenarios no cubiertos por LDA. Los modelos jerárquicos Raphil (referencia) y Pachinko Allocation Model (referencia) son un ejemplo de ello ya que diferencia de los modelos anteriores estos modelos rotornan una abstracción tópica multicapas con el objetivo de revelar las relaciones jerárqicas entre los tópicos. Otro escenario para el que creo un modelo basándose en las características de LDA fue el de los modelos dinámicos. Los modelos dinámicos que no aumen que los tópicos son estáticos sino que apuestan por el dinamismo de los temas latentes a través del tiempo. Este nuevo enfoque sirvió para realizar tareas donde el corpus contaba con elementos secuenciales como blogs, tweets o revistas. Topic Over Time (referencia) propuesto por Wang X, Mccallum A en 2006 fue uno de los primeros modelos desarrollados con este enfoque.

Si bien LDA es preciso y ampliamente utilizado, su enfoque presenta limitaciones al

trabajar con grandes volúmenes de datos debido al aumento en la cantidad de variables según el tamaño del corpus y el número de temas a identificar.

Para abordar las limitaciones de modelos clásicos como LDA, surgieron modelos tópicos neuronales, que adoptan redes neuronales profundas para modelar temas latentes. Estos modelos pueden inferir de manera eficiente los parámetros a través de la retropropagación automática del gradiente, permitiendo entrenar el modelo independientemente del tamaño del conjunto de datos, lo que les proporciona a los modelos tópicos neuronales mayor escalabilidad y flexibilidad. Esto dió lugar al surgimiento de varios modelos destinados a diversos escenarios.

Los modelos tópicos tienen aplicaciones en diversos escenarios. Algunos modelos, como LDA, funcionan bien para descubrir temas estáticos en colecciones de texto, pero en contextos como el de revistas científicas, donde los temas evolucionan con el tiempo, es necesario recurrir a modelos de temas dinámicos.

Los modelos de temas dinámicos permiten capturar la evolución de los temas a lo largo del tiempo, mostrando la aparición, modificación y desaparición de temas en una colección de documentos. Estos modelos son útiles en publicaciones periódicas o documentos que se generan secuencialmente, como revistas académicas o noticias. Los modelos dinámicos asumen que los documentos se dividen en intervalos de tiempo y cada intervalo contiene una distribución de temas que puede cambiar con el tiempo. Este enfoque permite representar la evolución de los temas, proporcionando una visión más precisa sobre los mismos.

En el ámbito académico cubano, las revistas científicas desempeñan un papel fundamental en la difusión del conocimiento y en la construcción de una comunidad académica sólida. Estas publicaciones periódicas permiten a los investigadores de cada especialidad compartir sus hallazgos y avances, generando un intercambio de información que facilita la actualización del conocimiento. A través de las revistas científicas, la investigación cubana puede ganar visibilidad y facilitar la creación de redes de colaboración entre científicos.

Estas revistas son el resultado de un proceso dinámico de investigación. Los temas que abordan evolucionan continuamente, reflejando tanto las tendencias globales como las necesidades específicas de la sociedad y el desarrollo científico local. Esta evolución de los temas permite observar las áreas prioritarias y emergentes en la ciencia. Así, analizar la evolución de los temas en las publicaciones científicas puede aportar información valiosa sobre el rumbo de la investigación en el país y ayudar a identificar áreas estratégicas para el desarrollo.

Las herramientas actuales de análisis de texto desarrolladas en nuestro país presentan limitaciones significativas cuando se trata de analizar la evolución de los temas, pues son implementaciones de modelos tópicos estáticos. A pesar de la necesidad existente, no se cuenta hasta el momento con una herramienta efectiva que permita analizar de manera sistemática y automatizada la evolución de los temas de investigación en las revistas científicas de Cuba. Aunque existe una colección determinada de artículos y

publicaciones disponibles, aún no se cuenta con sistemas que faciliten el seguimiento de cómo los temas y tendencias cambian a lo largo del tiempo.

Para los investigadores, este análisis es valioso porque permite identificar tendencias, áreas emergentes y temas que han perdido relevancia. Conocer la dinámica de los temas ayuda a enfocar mejor los esfuerzos de investigación, ya que los científicos pueden dirigir sus estudios hacia áreas de mayor impacto.

Desde el punto de vista de la planificación científica, disponer de un análisis sobre la evolución de temas de investigación permite mejorar la toma de decisiones. Los responsables de formular los planes para desarrollar las investigaciones pueden identificar las áreas prioritarias en las que deben enfocarse los recursos y fomentar colaboraciones entre instituciones que trabajan en temas afines. Esta información permite que las políticas científicas se alineen con las necesidades actuales y futuras de la sociedad, apoyando una distribución óptima de recursos.

En cuanto al desarrollo académico, el análisis de la evolución de temas en las revistas científicas, permite identificar campos con potencial de crecimiento. Las universidades e instituciones educativas pueden adaptar sus programas de formación, asegurando que los futuros investigadores estén preparados en áreas emergentes de investigación. Esto, a su vez, fortalece la formación de una comunidad académica competitiva y actualizada.

La investigación busca solucionar el problema del análisis de tópicos en contextos específicos mediante la implementación de un modelo tópico dinámico utilizando la herramienta *TopMost* y como fuente de datos los artículos de la revista **Ciencias Médicas**. El objetivo es explorar cómo el modelo seleccionado, se adapta a un dominio especializado como el de una publicación científica. Esto resulta novedoso, ya que los modelos tópicos dinámicos generalmente se han entrenado en conjuntos de datos amplios y variados, como redes sociales, artículos de noticias o bases de datos científicas de alcance general.

Los modelos dinámicos de temas muestran una variabilidad significativa en sus resultados al ser aplicados en diferentes dominios. Esta variabilidad se debe principalmente a que las características de cada conjunto de datos son únicas y dependen del contexto en el que fueron recolectados.

Con esta investigación se busca aplicar el modelo a un conjunto de datos que, aunque está limitado a una sola disciplina, podría ofrecer perspectivas novedosas sobre cómo emergen y evolucionan los temas en un contexto académico específico. La revista **Ciencias Médicas** abarca un área de conocimiento relativamente acotada, por lo que será interesante observar cómo el modelo capta la evolución de temas especializados en el ámbito de la medicina y si identifica patrones relevantes que puedan aplicarse a otras disciplinas.

Al realizar el entrenamiento con los datos de esta revista, se podrán realizar ajustes específicos en sus hiperparámetro, para maximizar su efectividad en este nuevo contexto. Realizar estos ajustes permitirá observar cómo el modelo puede identificar patrones que resalten las variaciones en los temas abordados por la revista a lo largo del tiempo.

Este proceso de optimización podría generar resultados que no solo beneficien el análisis temático en el contexto de la revista **Ciencias Matemáticas**, sino que también aporten ideas sobre la evolución de temas en revistas de disciplinas especializadas. De esta manera, este ajuste de parámetros contribuirá a desarrollar modelos más versátiles que puedan adaptarse mejor a conjuntos de datos específicos y ofrecer resultados que reflejen mejor las particularidades de esos dominios.

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.

Con este trabajo se pretende no solo validar la adaptabilidad del modelo a un nuevo contexto, sino también establecer un precedente para futuras aplicaciones de modelos dinámicos en áreas de conocimiento que poseen características únicas y específicas. En última instancia, este enfoque podría facilitar la comprensión de la evolución de temas en dominios especializados y brindar una herramienta más robusta para el análisis académico en diversas áreas del conocimiento.