El modelado de tópicos y SVD

Antes de explicar qué es el modelado de tópicos primero respondamos la pregunta ¿Qué es un tópico? En lingüística, un tópico puede definirse como la idea principal de un texto. Puede presentarse a diferentes granularidades. Por ejemplo, se puede hablar del tópico de una sentencia, de un párrafo, de un artículo o de un conjunto de documentos, entre otros. El descubrimiento y análisis de tópicos tiene aplicación en diversos ámbitos, entre ellos, los estudios de opinión pública y el marketing publicitario y político, por ejemplo, saber:

- De qué están hablando los usuarios de una red social en un momento determinado.
- Qué líneas de investigación prevalecen en un área de estudio determinada y cómo han cambiado en el tiempo.
- Qué es lo que gusta de un producto, lo cual requiere descubrir tópicos tanto en revisiones positivas como negativas de los usuarios.
- Cuál es la opinión de una comunidad frente a políticas gubernamentales.
- Cuáles son los temas principales que se debaten en una elección presidencial.

El modelado de tópicos *(topic modeling*) es una de las tareas del procesamiento del lenguaje natural (PLN) que se utiliza para identificar los temas clave presentes en un conjunto de documentos. El objetivo principal es descubrir la estructura subyacente que estos abordan sin necesidad de etiquetas previas o supervisión.

Un modelo de tópicos asume que cada documento es una mezcla de varios temas y que cada palabra en el documento contribuye a estos de alguna manera. Así, el modelado de tópicos permite obtener los aspectos semánticos clave en los textos, mediante el descubrimiento de patrones de uso de palabras y cómo estos conectan documentos que comparten regularidades similares. Estos patrones recurrentes representan los tópicos: conjuntos de palabras que tienden a aparecer juntas en los mismos contextos, por lo que se asume que hacen referencia a los mismos temas.

Además de los textos, también es posible disponer de otra información que puede ser utilizada como contexto adicional para analizar los tópicos. Por ejemplo, el momento o lugar en el que los

textos fueron generados, sus autores y sus fuentes. Esta metadata (variables de contexto) podría estar asociada a los tópicos que se descubran y ayudar en su análisis. Por ejemplo, analizar los tópicos en el tiempo permitiría descubrir si hay una tendencia (*trending topic*) o si algún tema está dejando de ser interesante para los usuarios. También, examinar tópicos en diferentes lugares podría proporcionar conocimiento acerca de cómo se matizan o varían las opiniones de las personas según su ubicación.

¿Cómo descubrir tópicos en colecciones de documentos? En la representación de bolsa de palabras (*Bag of Words*, BOW), de alta dimensionalidad, considerar los términos de manera individual tiene limitaciones debido a que múltiples documentos pueden discutir las mismas ideas utilizando palabras diferentes (el mismo concepto puede ser expresado usando cualquier número de términos) y el mismo término puede tener varios significados en diferentes contextos. Pero si se aplica una reducción de la dimensionalidad, es posible obtener un nuevo espacio en el cual se "reducen" términos que tienen la misma semántica (están correlacionados), y se identifican y clarifican términos con múltiples significados, lo cual facilita la comprensión de cada uno de estos dependiendo del contexto. Además, se obtiene una representación de baja dimensión de los documentos que refleja conceptos en lugar de términos aislados. Este nuevo espacio de baja dimensión se conoce como **espacio semántico latente**.

La premisa es que, al aplicar una reducción de la dimensionalidad a la matriz de términos-documentos, se deriva una nueva representación que revela la estructura de tópicos del corpus más claramente que la representación original. La reducción de la dimensionalidad es capaz de llevar la representación BOW a un nivel más abstracto, en el cual las nuevas dimensiones corresponden a conceptos o tópicos. De esta manera, diferentes formas de expresar el mismo contenido pueden ser reducido a una representación común y términos con múltiples significados pueden ser identificados.

En general, los algoritmos de modelado de tópicos son métodos para la reducción de la dimensionalidad que identifican las relaciones de términos y documentos a partir de las dimensiones obtenidas para representar el espacio semántico latente. Entre estos, podemos mencionar la Indexación semántica latente, que explicaremos a continuación.

Indexación semántica latente

La Indexación Semántica Latente (LSI), también conocida como Análisis Semántico Latente (LSA), se basa en una descomposición en valores singulares (SVD) de la matriz de términos-documentos, la cual construye una aproximación de bajo rango de la matriz original (SVD reducida), preservando la similitud entre documentos. Es decir, LSI proyecta tanto documentos como palabras en un espacio latente $k - dimensional$. Estas nuevas dimensiones se interpretan como conceptos semánticos. Es posible entonces el análisis de documentos a nivel conceptual, ya que se crean asociaciones entre textos temáticamente relacionados. Además, análogo a la similitud de documentos, la similitud de términos puede ser medida en el espacio semántico latente para identificar términos con significados similares.

A continuación, se describe cómo se puede utilizar SVD para construir un modelo de tópicos.

- Representación de Documentos. El proceso comienza con una matriz término-documento ($\boldsymbol{X}$), donde las filas representan palabras y las columnas representan documentos. Cada entrada de la matriz contiene la frecuencia de un término en un documento. Una de las técnicas más utilizada para realizar esta ponderación es TF-IDF (*Term Frequency-Inverse Document Frequency*).
- Descomposición en valores singulares (SVD). A continuación, se aplica la descomposición en valores singulares a esta matriz: $\boldsymbol{X = U\Sigma V'}$.
- Reducción de Dimensionalidad. Tras la descomposición, se puede reducir la dimensionalidad truncando algunas de las dimensiones menos importantes. Esto se logra manteniendo solo los primeros k valores singulares: $\boldsymbol{X_k = U_k\Sigma_k V'_k}$. Estas dimensiones son los campos semánticos.
- Matrices Resultantes. Las matrices $\boldsymbol{U_k}$ y $\boldsymbol{V_k}$ reducidas se utilizan como representación latente de términos y documentos, respectivamente. La matriz diagonal $\boldsymbol{\Sigma_k}$ contiene los valores singulares, que proporcionan información sobre la importancia relativa de las dimensiones.
- Identificación de Tópicos. Los vectores en $\boldsymbol{U_k}$ y $\boldsymbol{V_k}$ se pueden utilizar para identificar los tópicos latentes en los documentos y términos. $\boldsymbol{U_k}$ es la matriz término-tópico, en la que cada fila representa un término y cada columna un tópico. Por su parte $\boldsymbol{V_k}$ representa la relación entre documentos y los tópicos. Así, después de aplicar la SVD reducida, cada término tendrá una

"posibilidad" de estar en un documento. De la misma forma, cada documento tendrá una "posibilidad" de contener esos términos (ver Fig. 1).

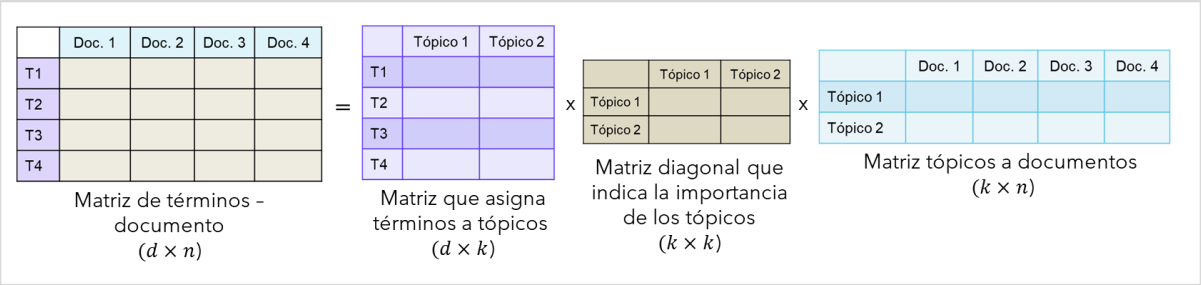

Fig. 1. La matriz resultante de la descomposición en valores singulares se puede utilizar para identificar los tópicos principales en los documentos. Los vectores singulares derechos representan los tópicos, mientras que los valores singulares representan la importancia relativa de cada tópico. Los vectores singulares izquierdos se pueden utilizar para encontrar las palabras más importantes en cada tópico.

Es importante resaltar que documentos que tratan un tópico particular (comparten un campo semántico) serán más similares en esta representación, aun si no comparten todas las palabras. Así, la similitud entre dos documentos puede determinarse en este espacio semántico, la cual puede ser utilizada para diferentes tipos de análisis. Pero también será posible calcular la similitud entre términos, por ejemplo, para determinar palabras con significados similares de acuerdo con el contexto.

¿Cómo se interpreta un modelo de tópicos? La manera común de interpretar modelos de tópicos es a través de la inspección de las asociaciones entre los términos y los temas. ¿Cómo? Se examinan los primeros términos que están fuertemente asociados con cada tópico, así como la contribución de cada tópico a cada documento. Por ejemplo, para LSI, los términos están ordenados de acuerdo con el coeficiente correspondiente a una característica dada en el espacio semántico. Acá, la representación original de las características juega un rol clave en la definición de los tópicos y en su identificación dentro cada documento. Como resultado, se obtiene una representación entendible de documentos que es útil para analizar los temas presentes en estos.

Es importante destacar que, si bien LSI utilizando SVD es un enfoque valioso, ha sido superado en algunos aspectos por modelos más avanzados de tópicos, como *Latent Dirichlet Allocation* (LDA), el cual proporciona una base probabilística para reducción de la dimensionalidad. Este método permite razonar acerca de los tópicos presentes en un documento expresando la probabilidad de

considerar cada palabra en cualquier tópico dado, y cada documento se asigna a los tópicos con diferentes pesos. De esta manera, se facilita la interpretación del significado de los tópicos. Sin embargo, la aplicación de SVD para modelar tópicos sigue siendo una técnica útil en ciertos contextos y puede proporcionar *insights* valiosos en el análisis de textos.

Bibliografía

Albrecht, J., Ramachandran, S., Winkler, C. (2020). Blueprints for Text Analytics Using Python. O'Reilly Media, Inc.

"Singular Value Decomposition: Image Processing, Natural Language Processing, and Social Media" En: Nelson, H. (2023). Essential Math for AI. Capítulo 6. O'Reilly Media, Inc.

Universidad de los Andes | Vigilada Mineducación Reconocimiento como
Universidad: Decreto 1297 del 30 de mayo de 1964.
Reconocimiento personería jurídica: Resolución 28 del 23 de febrero de 1949 Minjusticia.