Maestría en Inteligencia Artificial Aplicada (MNA)

Modelado de Tópicos: LSI

Procesamiento de Lenguaje Natural (NLP)

Luis Eduardo Falcón Morales

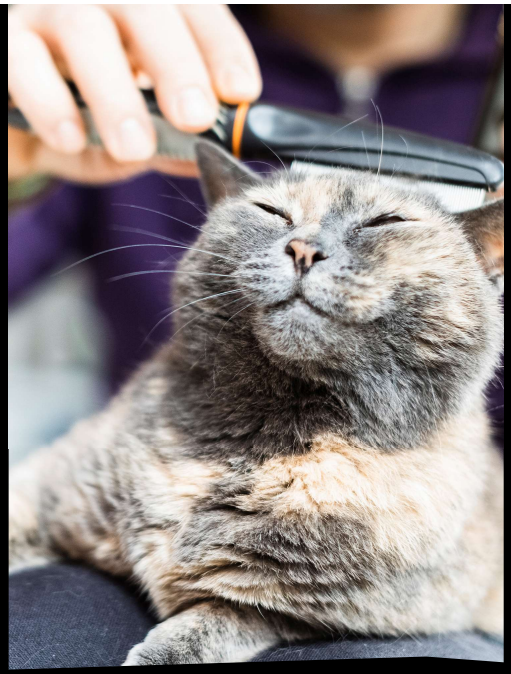
Modelado de Tópicos : Topic Modeling

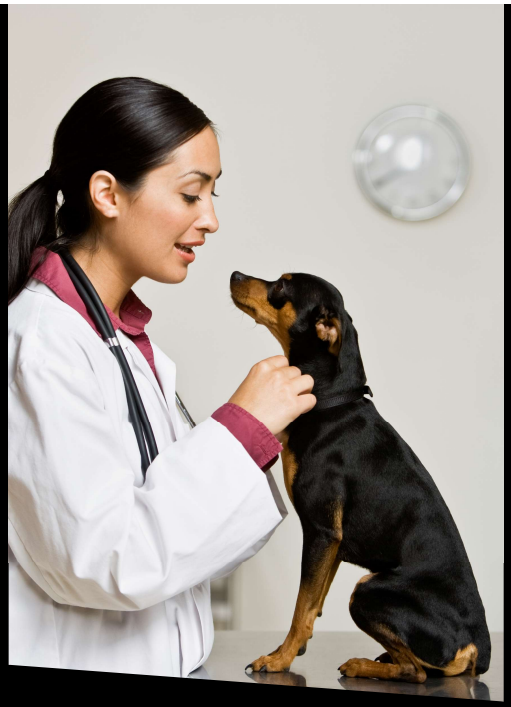
A medida que los recursos computacionales empiezan a ser más poderosos, las bases de datos más grandes y los modelos más robustos, los problemas que se abordan también empiezan a ser más complejos.

Esto nos lleva a tener más información de la que se pueda procesar y analizar. Se genera tanta información en tan poco tiempo que mucha de esta información se queda sin etiquetar y por lo tanto sin poder usarse con las técnicas de aprendizaje supervisado. Así, las técnicas estudiadas en modelado de tópicos (topic modeling en inglés) se consideran dentro del área de aprendizaje no supervisado que estudiaremos en esta semana.

- Debido a la gran cantidad de información que ya se genera en cualquier departamento de una empresa, se requieren técnicas que nos ayuden a organizar, buscar y entender todos estos documentos.

- El tipo de datos de los que estamos hablando pueden ser tanto estructurados, como no estructurados. Y en particular de los no estructurados, los relacionados con documentos de texto son en promedio más del 30% y puede incrementarse rápidamente de acuerdo al giro de la empresa.





- Por ejemplo, a los diseñadores del sitio web de una empresa junto con el equipo de mercadotecnia, les interesa aumentar el tráfico de su página, a la vez que reclutan una mayor cantidad de clientes potenciales. Esta área se conoce usualmente como “optimización en los motores de búsqueda”, SEO, por sus siglas en inglés (Search Engine Optimization).

- Cerca del 70% de una experiencia en línea inicia en un motor de búsqueda: ¿qué palabras claves son las que utilizan las personas para buscar servicios o productos para, por ejemplo, sus bebés o mascotas?

- Conocer estas palabras clave ayudará a posicionar mejor tu sitio web y a conocer mejor a tus futuros clientes.

Ya hemos estudiado la nube de palabras (BOW bag-of-words) y nos ayuda a identificar las palabras clave por clases en un problema de análisis de sentimiento.

Sin embargo, esto no es suficiente porque no nos ayuda a identificar la manera en que una palabra dada participa dentro del tema de un documento.

- Los **bancos** han implementado políticas más agresivas para recuperar carteras vencidas a partir de este año.
- Estudiaremos la dinámica en el agua de la trayectoria de los **bancos** de peces mediante ecuaciones diferenciales.
- El parque de la colonia que recién inauguró el presidente municipal tiene unos **bancos** para descansar con unos diseños de personajes animados que han gustado mucho a los niños y niñas.

Modelado de temas como métodos no supervisados

El área conocida como **Modelado de Temas o Tópicos** (**Topic Modeling** en inglés) nos proporciona diversos métodos para organizar, buscar entender y resumir grandes cantidades de documentos de texto.

La idea principal de todos estos métodos de modelado de temas, es que consideran cada documento como una colección de temas (topics), en cierta proporción. Y a su vez, cada tema es una colección de palabras clave (keywords), también en cierta proporción.

A través de la búsqueda de la distribución de los temas en cada documento y la distribución de las palabras clave en cada tema, es que estos métodos de modelado de temas tratan de organizar, buscar y entender cada documento.

En general, estos métodos de modelado de temas se consideran como no supervisados, ya que la extracción de dicha información se hará de esta manera no supervisada.

Cuando hablamos del “tema” (topic) de un documento, nos referimos al conjunto de palabras clave (keywords) que son representativas del contenido de dicho documento.

Así, si identificamos dichas palabras clave en cualquier documento dado, sabremos cuál será el tema sobre el cual se está hablando.

Uno de los principales retos que se tienen en este tipo de problemas, es que en general no se conoce de entrada la cantidad de temas que existen en todos los documentos. Sin embargo, dependiendo del tipo de negocio, el número de temas a veces se puede aproximar. En el caso de una agencia de noticias, por ejemplo, por experiencia se sabe la manera en que estas se clasifican: política, cultura, espectáculos, deportes, sociales, entre otras, todas ellas a su vez en local, nacional e internacional.

Existe una gran cantidad de métodos de modelado de temas (topic modeling) y más que querer ver una gran cantidad de ellos, estudiaremos dos de los más conocidos y que nos ayuden a entender el concepto y la manera en que pueden aplicarse en diferentes situaciones reales.

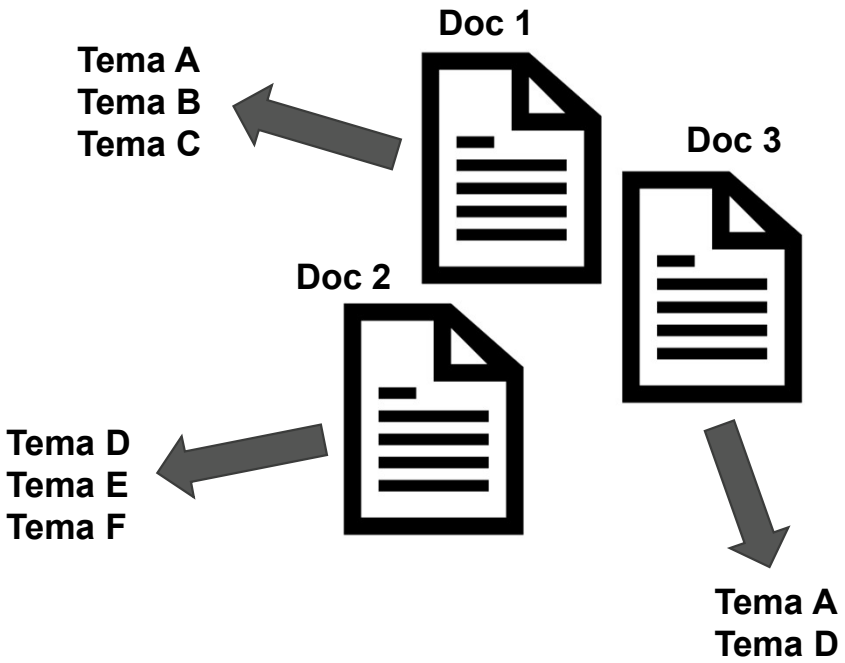
En particular estudiaremos los dos siguientes:

- Indexación Semántica Latente (LSI – Latent Semantic Indexing)
- Asignación Latente de Dirichlet (LDA - Latent Dirichlet Allocation)

Análisis Semántico Latente / Indexación Semántica Latente

[Latent Semantic Analysis (LSA) / Latent Semantic Indexing (LSI)]

Supongamos que cada documento está formado por un conjunto de diversos temas.



Y a su vez cada tema está formado por un conjunto de palabras clave.

Tema A : {palabras clave de A}
Tema B : {palabras clave de B}
Tema C : {palabras clave de C}
Tema D : {palabras clave de D}
Tema E : {palabras clave de E}
Tema F : {palabras clave de F}

El objetivo es encontrar la proporción o distribución de cada tema en cada documento, así como la distribución de cada conjunto de palabras clave de cada tema.

Document Term/Word Matrix

	Term_1	Term_2	Term_3	…	Term_n
Doc_1	0	1.67	0.53	⋯	0
Doc_2	0	0.87	0.31	⋯	0
Doc_3	1.87	0	0	⋯	2.53
⋮	⋮	⋮	⋮	⋱	⋮
Doc_m	0.33	0	1.57	⋯	0

Token ≈ Term ≈ Word
Documento ≈ artículo
Topic ≈ Tópico ≈ Tema

Cada valor numérico en la tabla indica por ejemplo el valor tf-idf de una Document-Term matrix, es decir, son valores que nos hablan de la importancia cada palabra o término dentro de cada documento.

¿Podríamos construir ahora, a partir de estas dos marices, una matriz que nos indique qué palabras (Terms) son las que aparecen en cada Tema?

Document Topic/Tema Matrix

	Tema_1	Tema_2	Tema_3	…	Tema_r
Doc_1	0.15	0.34	0.22	⋯	0.05
Doc_2	0	0.66	0.14	⋯	0.10
Doc_3	0.35	0	0.36	⋯	0.07
⋮	⋮	⋮	⋮	⋱	⋮
Doc_m	0.01	0.08	0.61	⋯	0.17

Un documento puede abordar varios temas/tópicos. Así, cada valor numérico en la tabla nos indica que tan viable es que el tema de dicha columna se trate en el documento de dicho renglón. Se le suele llamar Document-Topic matrix, o bien Document-Latent-Concept matrix.

Modelado de temas y matriz de valores singulares (SVD):

Por lo tanto, ahora se desea construir una matriz que nos indique qué palabras/tokens/términos son los que usualmente se utilizan cuando se habla de cierto tema. Es decir, deseamos encontrar las palabras clave (keywords) de cada tema.

Term - Topic Matrix

	Tema_1	Tema_2	Tema_3	…	Tema_r
Term_1		?		…	
Term_2				…	
Term_3				…	
⋮	⋮	⋮	⋮	⋱	⋮
Term_n				…	

Veremos que una manera de construir esta matriz es a través de la llamada Descomposición en Valores Singulares, SVD, por sus siglas en inglés (Singular Value Decomposition).

Este método funciona además como una técnica de reducción de dimensionalidad, ya que queremos encontrar la cantidad mínima de palabras (Terms) que describa cada Tema.

< Terms (palabras) : r >

< Documentos : m >

	Te1	Te2	Te3		Tek	
D1	a11	a12	a13			
D2	a21	a22	a23			
D3	a31	a32	a33			
Dj					ajk	

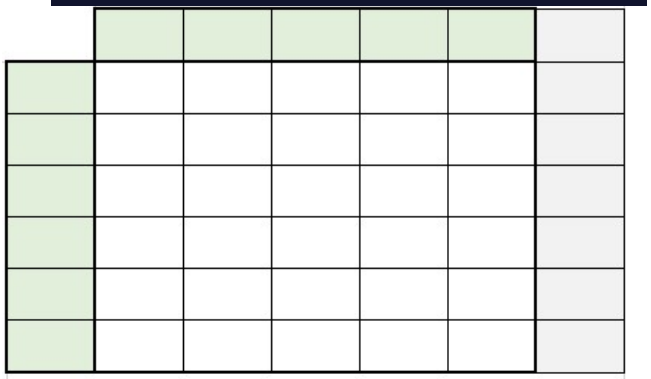
Document Term Matrix (DTM)

Token ≈ Term ≈ Word
Documento ≈ artículo

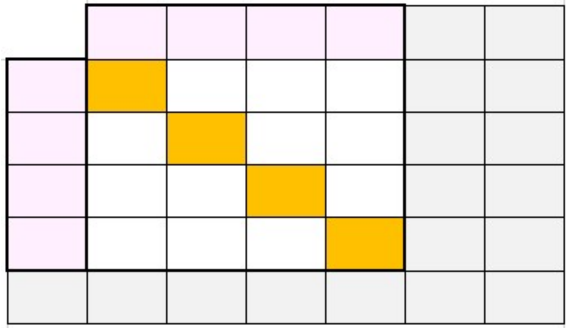
Descomposición SVD:
DTM → SVD → $U * D * V^T$

r Conceptos Latentes

Matriz de Variables/Conceptos Latentes de los Documentos: Similaridad entre Documentos

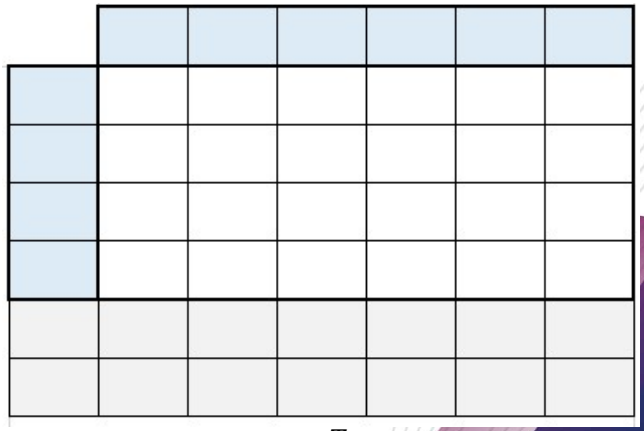


$U : m \times r$



$D : r \times r$

Matriz de Variables/Conceptos Latentes de los Términos (Palabras): Relación entre par de palabras en un Documento



$V^T : r \times n$

< Terms (palabras) : r >

< Documentos : m >

	Te1	Te2	Te3		Tek	
D1	a11	a12	a13			
D2	a21	a22	a23			
D3	a31	a32	a33			
Dj					ajk	

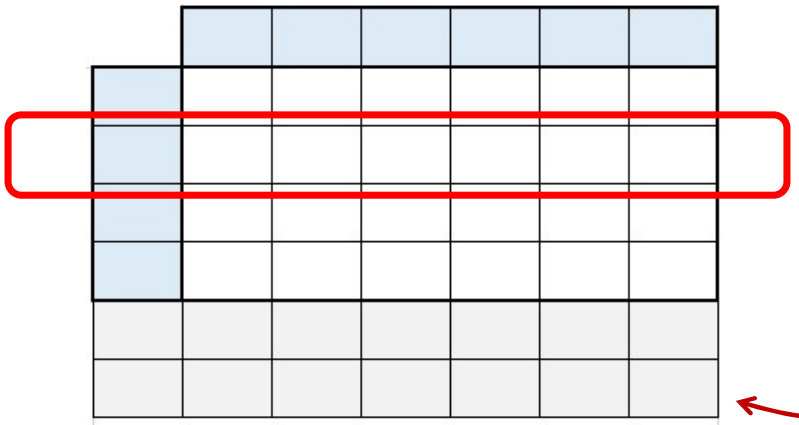
Document Term Matrix (DTM)

Descomposición SVD:
DTM → SVD → $U * D * V^T$

Reducción de Dimensionalidad

***r* Conceptos Latentes**

Matriz de Variables/Conceptos Latentes de los Documentos



$V^T : r \times n$

En esta matriz, la información de las palabras queda integrada dentro de los valores de dicha matriz y sustituida por los r conceptos latentes, que estarán explicando/expresando de manera más compacta los documentos.

En resumen, ¿qué entendemos por "Topic/Tópico" en esta técnica no supervisada? Por "Topic/Tópico" entenderemos la información que nos regresan los renglones de la matriz V^T al aplicar la SVD.

Así, las palabras que aparecen en cada uno de los "r" renglones, con sus pesos respectivos, serán los "r" Topics/Tópicos que identifica este método de LSI.

Las columnas son las N palabras que definiste en tu vocabulario.

$V^T : r \times n$

	Word1	Word2	Word3	Word4		WordN
Topic-1						
Topic-2						
Topic-r						

El método de indexación semántica latente (LSI) se basa en esta descomposición, donde cada documento queda expresado por una cantidad de conceptos/tópicos latentes, dados a su vez por un conjunto de palabras clave:

Teorema: Descomposición mediante SVD

Dada cualquier matriz de valores reales $A_{m\times n}$, existe una descomposición SVD de la forma:

$$A = U\Sigma V^T$$

donde $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$, y $\sigma_{r+1} = \cdots = \sigma_n = 0$.

Entonces:

$$A = \sigma_1 u_1 v_1^T + \sigma_2 u_2 v_2^T + \cdots + \sigma_r u_r v_r^T$$

De esta descomposición podemos aproximar la matriz A con respecto a los $d \leq r$ sumandos más grandes. Así, la cantidad de información contenida en dicha aproximación estaría dada como:

$$\frac{\sum_{k=1}^{d} \sigma_k}{\sum_{k=1}^{r} \sigma_k}$$

Y por lo tanto la pérdida de información con dicha aproximación sería:

$$1 - \frac{\sum_{k=1}^{d} \sigma_k}{\sum_{k=1}^{r} \sigma_k}$$

Tecnológico de Monterrey

D.R.© Tecnológico de Monterrey, México, 2022.
Prohibida la reproducción total o parcial
de esta obra sin expresa autorización del
Tecnológico de Monterrey.