



Maestría en Inteligencia Artificial Aplicada (MNA)

# Modelado de Tópicos: LDA

Procesamiento de Lenguaje Natural (NLP)

Luis Eduardo Falcón Morales

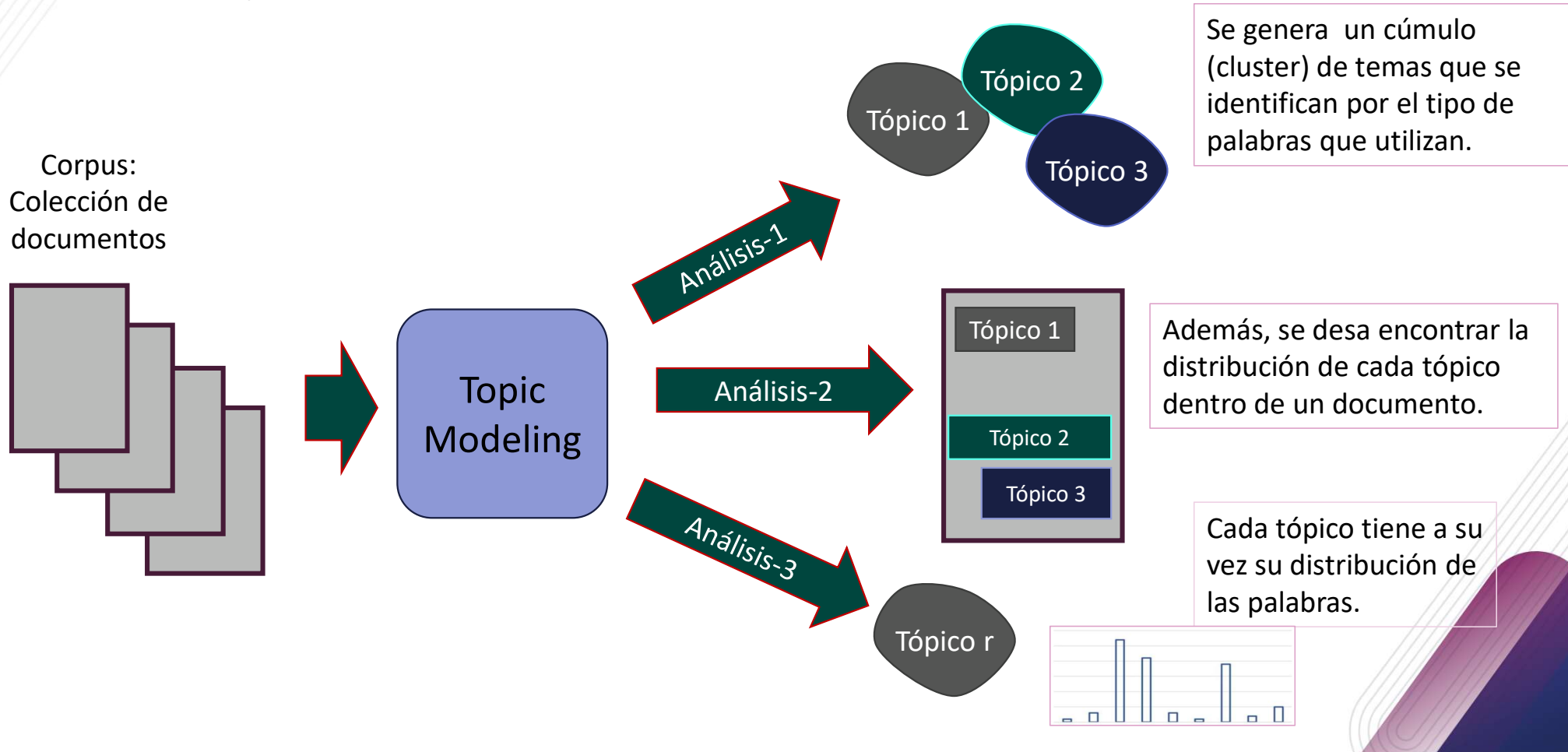


# Asignación Latente de Dirichlet

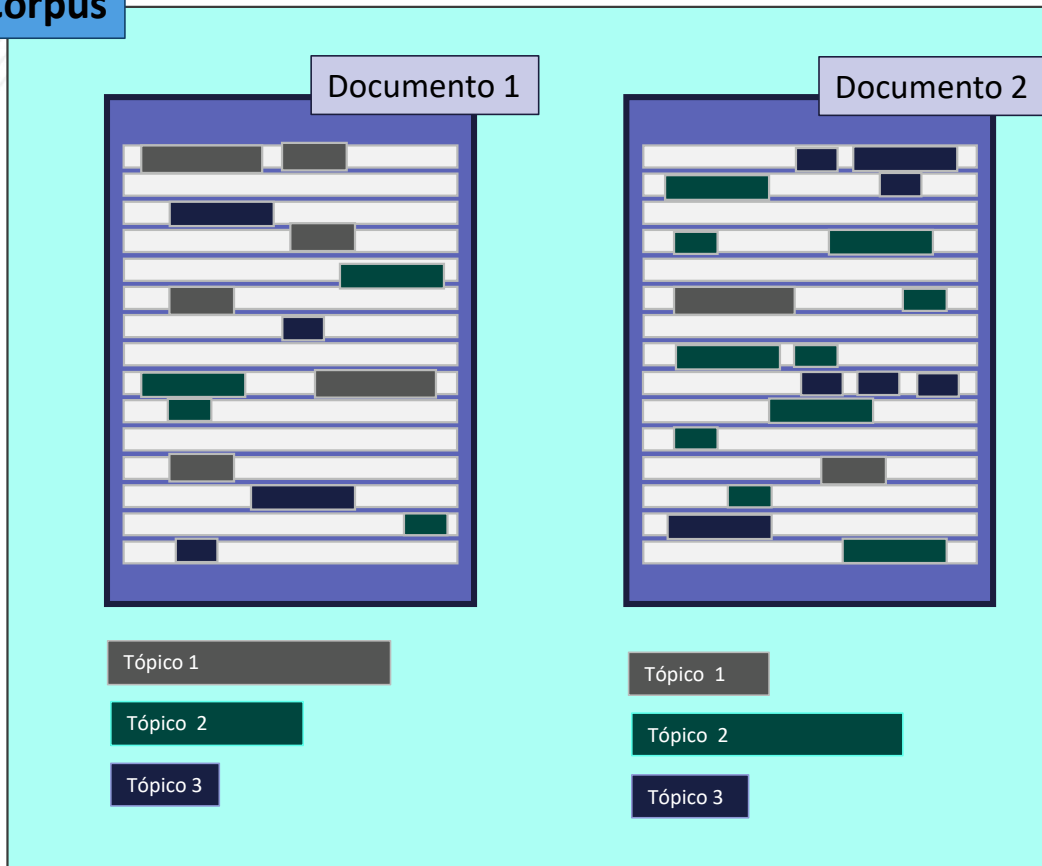
## [ Latent Dirichlet Allocation (LDA) ]



La técnica LDA está dentro de área de modelado de temas, técnica no supervisada para encontrar diferentes temas dentro de un documento de texto y entre diferentes documentos entre sí.



## Corpus



Un Documento puede ser un artículo, un libro, un tweet, un email, una oración, etc.

La técnica LDA se basa en modelos probabilísticos de las distribuciones de los tópicos dentro de cada documento y de la distribución de los tokens dentro de cada tópico. Por ello, existen dos hiperparámetros principalmente en esta técnica llamados alfa y beta.

Alfa representa el valor de densidad de la relación documento-tópico y Beta representa la densidad de la relación tópico-token.

Tópico 1

Tópico 2

Tópico 3

Divisas  
Banco  
Depósito  
...

Libro  
Ficción  
Novela  
...

Ecuación  
Variable  
Integral  
...



## ¿Cómo procede el modelado de tópicos?

D1 : Me gustan las nueces y las manzanas.

D2 : Desayuné cereal con leche.

D3 : Los perros y los gatos son buenas mascotas.

D4 : Ayer, mi vecino adoptó un perrito.

D5 : Al perico del vecino le dan de comer manzanas.

---

Tópico A: 30% manzanas, 20% nueces, 10% cereal, 8% desayuno, 5% leche,...

Tópico B : 20% perros, 20% gatos, 25% mascotas, 15% perico, 5% adopción, 5% vecino...

---

D1 y D2 : 100% Tópico A

D3 y D4 : 100% Tópico B

D5 : 60% Tópico A y 40% Tópico B

---



Un tópico es una combinación de palabras con una distribución de probabilidad.

Un documento es una combinación de tópicos con otra distribución de probabilidad.



## Document Term Matrix

	word1	word2	...	wordM
Doc1				
Doc2				
Doc3				
...				
DocN				

Factorización  
LDA



La técnica LDA también genera una factorización en dos matrices a partir de la document-term matrix: la matriz Documento-Tópico y la matriz Tópico-Tokens.

Sin embargo, la manera en que se genera se basa en modelar cada una de las distribuciones probabilísticas de ambas relaciones. No entraremos en los detalles matemáticos de esta solución, pero puedes revisar la bibliografía correspondiente para abundar en los detalles.

	topic1	...	topicK
Doc1			
Doc2			
Doc3			
...			
DocN			

\*

	word1	word2	...	wordM
topic1				
...				
topicK				

- 
- El método Latent Dirichlet Allocation (LDA) es de los algoritmos más utilizados dentro del área de modelado de temas (topic modeling).
  - LDA considera cada documento como una combinación de temas y cada tema a su vez, lo considera como una combinación de palabras clave.
  - Esta técnica no busca separar o asignar a un documento un solo tópico, sino que se permite que exista traslapa de temas dentro de un mismo documento.
  - LDA es una técnica matemática que nos permite atacar ambos problemas en un mismo procedimiento.
- 



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