

# Topic Models

9/

## LATENT Dirichlet Allocation (LDA)

Inicio con LCA para buscar dentro de las dimensiones la esencia de un Topico

→ Probabilístico → LDA

### Claves

- Descubrir la temática oculta en grandes cantidades de documentos
- Anotar documentos
- Utilizar las anotaciones para visualizar, organizar, resumir, etc.

Viendo las palabras principales

- ? ¿Descubrir de que habla el doc?
- ¿Un solo tema o varios?
- ¿Algún tema sobresale?
- ¿Es necesario leerlo en orden?

Topico: distribución probabilística sobre todo el vocabulario

Topico 1	
Jugador	0.098
partido	0.095
pelota	0.087
⋮	

Buscar cuáles son los tópicos presentes en mi corpus

LDA → Proceso  
→ Generativo  
→ Modelo Probabilístico



→ Todos los tópicos tienen TODAS las palabras pero con distintas probabilidades

Algoritmo GENERATIVO (SUPONE los tópicos <sup>como</sup> DADOS)

1) Crea un Documento VACÍO

2) ASIGNAMOS AL Documento, una proporción (AZAR)  
de los distintos tópicos que tenemos.

3) GENERAMOS palabras de modo tal:

- Elegimos tópico al AZAR (teniendo en cuenta las proporciones del iter 2)
- Dado el tópico, elegir una palabra según las probabilidades en ese tópico.
- Repetimos desde (3) hasta generar "N" palabras.

4) Repetir desde (1) hasta generar D documentos.

Modelo LDA Gráfico también aparece

Distribución de  
Dirichlet( $\eta$ )

→ Busca que en un documento hay una  
cantidad "Acotada" de tópicos.

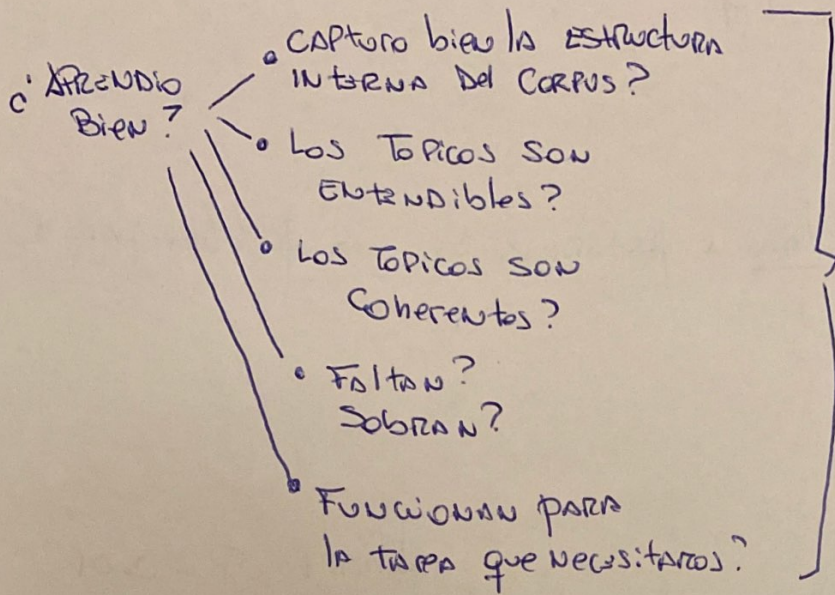
"Cuántos tópicos voy a tener activos por documento"

[ Como hallamos la distribución de probabilidades de cada tópico  
" " " " De cada palabra en cada tópico ]

→ Herramientas Bayesianas

- Existen tópicos pre-entrenados.
- Buscar tópicos en Mi CORPUS.





## Métodos:

- Oxiómetro
- Intrínsecos
- Métricas basadas en Human Judgments
- Extrínsecos (evaluar sobre una tarea)

- Oxiómetro: Miro x Aseguro con el conocimiento del problema que tengo.

## MEDIDAS

Intrínsecas: USA Topic-Coherence, uso el PIG x otro CORPUS PARA COMPARAR.

## Human

Judgments: → Word Intrusion: Meto una Palabra con poca proba en ese tópico x Alta en otro. Luego veo si puedo identificar via "humano" si encuentro esa palabra intrusa

→ Topic Intrusion: Agrego un Tópico Intruso

Word Relevance: LDD's

Frecuencia de la palabra en el Corpus

Frecuencia estimada de la palabra en el tópico.

Ejemplo: Consumo Diferente de Revistas según sexo