Topic Models

LATENT Dirichlet Allocation (LDA)

Inicio con LCA para buscar dentro de las dimensiones la esencia de un Topico
↳ Probabilistico → LDA

Claves

- Descubrir la tematica 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 sob tema o varios?
¿Algun tema sobresale?
¿Es necesario leerlo en orden?

Topico: Distribución probabilidad sobre todo el vocabulario

Topico 1	
Jugador	0.098
Partido	0.095
Pelota	0.087
⋮	

Buscar cuales son los topicos presentes en mi corpus

LDA → Proceso
Generativo
Modelo Probabilistico

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

Algoritmo generativo (supone los tópicos como 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 item 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(η) ⇝ Busca que en un documento hay una cantidad "acotada" de tópicos.
"Cuantos 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.

¿Aprendió Bien?

- Capturo bien la estructura interna del corpus?
- Los tópicos son entendibles?
- Los tópicos son coherentes?
- Faltan? Sobran?
- Funcionan para la tarea que necesitamos?

Métodos:

- Ojímetro
- Intrínsecos
- Métricas basadas en Human Judgements
- Extrínsecos (evaluar sobre una tarea)

-**Ojímetro**: Mido y Asigno con el conocimiento del problema que tengo.

-**Medidas Intrínsecas**: Usa Topic-Coherance, uso el PMI y otro corpus para comparar.

-**Human Judgements**: → Word Intrusion: Meto una palabra con poca prob en ese tópico y alta en otro. Luego veo si puedo identificar vía "humano" si encuentro esa palabra intrusa

→ Topic Intrusion: Agrego un tópico intruso

Word Relevance: LDAvis

- Frecuencia de la palabra en el corpus
- Frecuencia estimada de la palabra en el tópico.

Ejemplo: Consumo Diferente de Revistas según sexo