M5. Minería de Texto + webscraping

Clase 4. Modelado de tópicos 1. LDA

factor~data
IDAES_UNSAM

¿Cómo representar “matemáticamente” un texto?

Tidy Text

No es la conciencia (...) la que determina su ser sino (...) el ser social lo que determina su conciencia.

doc	word
1	no
1	es
1	la
1	conciencia
1|	la
1	que
1	determina
1	su
1	ser
...	...

Tidy Text

Un fantasma recorre Europa: el fantasma del comunismo

doc	word
2	un
2	fantasma
2	recorre
2	europa
2	el
2	fantasma
2	del
2	comunismo

Tidy Text

No es la conciencia (...) la que determina su ser sino (...) el ser social lo que determina su conciencia.

Un fantasma recorre Europa: el fantasma del comunismo

doc	word
1	no
1	es
1	la
...	...
2	el
2	fantasma
2	del
2	comunismo

Tidy Text

doc	word
1	no
1	es
1	la
...	...
2	el
2	fantasma
2	del
2	comunismo

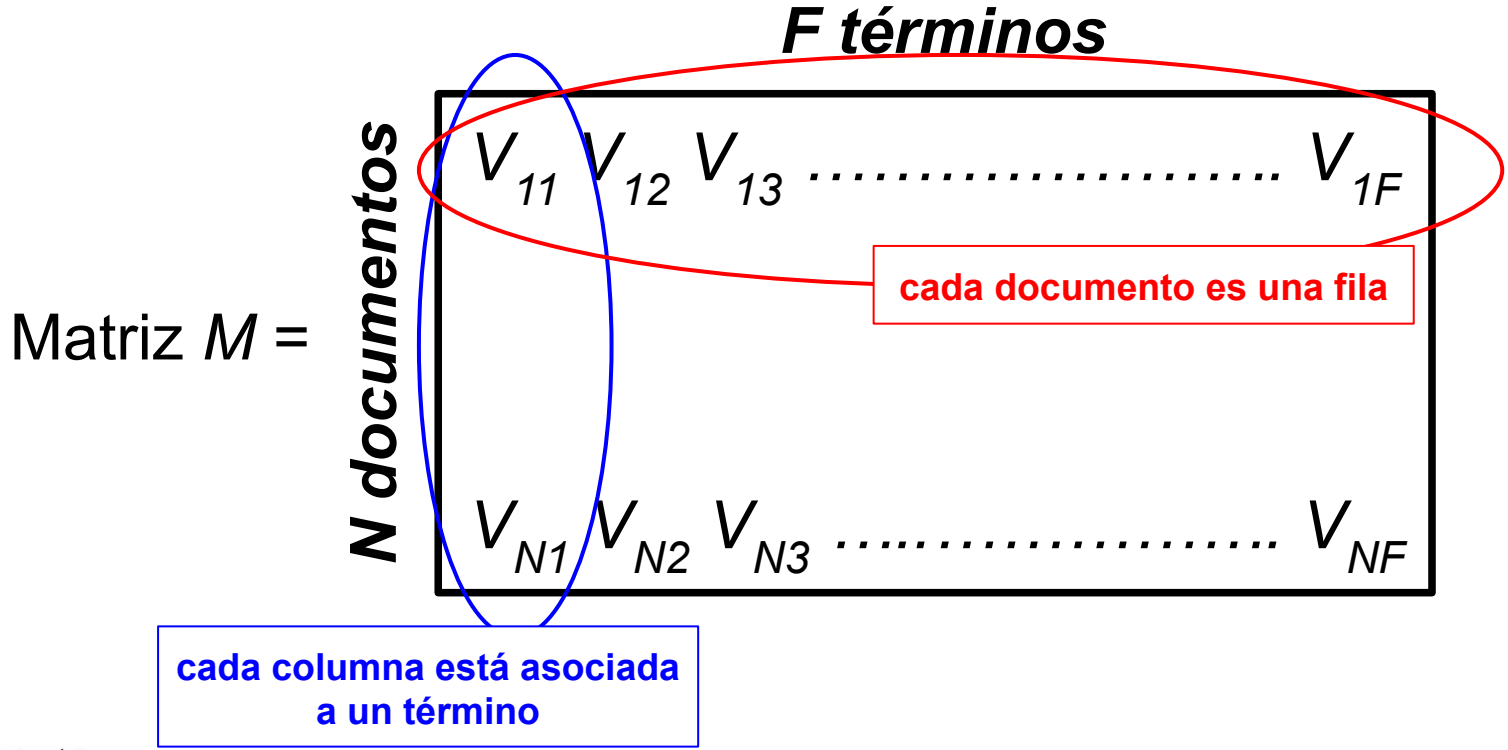
```
group_by(doc) %>%
    count(word)
```

doc	word	count
1	no	1
1	es	1
1	la	2
1	conciencia	2
...	...	...
2	el	1
2	fantasma	2
2	del	1
2	comunismo	1

Document-Term Matrix (TFM)

doc	no	es	la	conciencia	...	el	fantasma	del	comunismo
1	1	1	2	2	...	0	0	0	0
2	0	1	0	0	...	1	2	1	1

Document-Term Matrix (TFM)

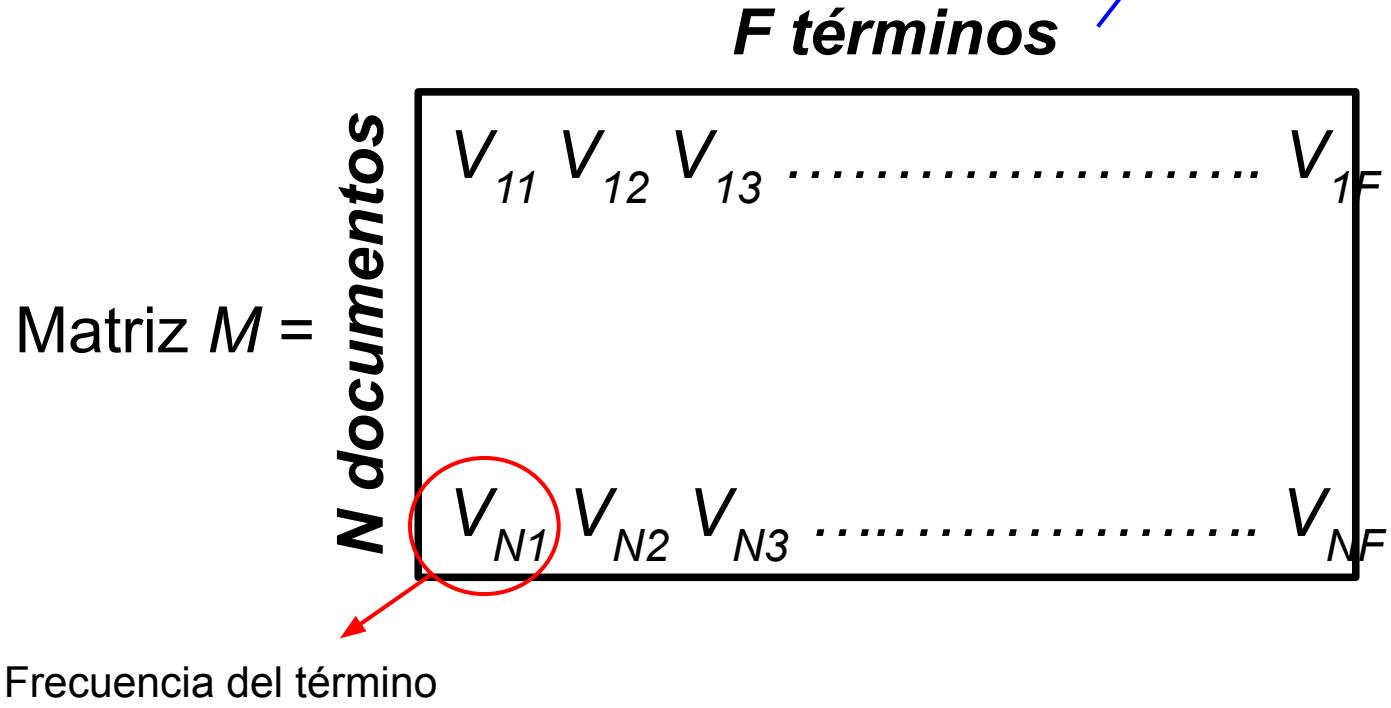


factor~data
IDAES_UNSAM

Fuente: https://github.com/gefero/ws_text_mining/blob/master/slides/04_pinto.pdf

Document-Term Matrix (TFM)

Palabras, bigramas, trigramas, lemas, solo la raíz de la palabra...



Fuente: https://github.com/gefero/ws_text_mining/blob/master/slides/04_pinto.pdf

Bag of Words (BoW)

- Representación de cada documento en función de las palabras que contiene

- Características:
  - Es simple de generar
  - Se asume que las palabras son “independientes”
  - Los vectores son claramente no independientes
  - La gramática y el orden de las palabras se pierden

Modelado de tópicos

¿Qué es?

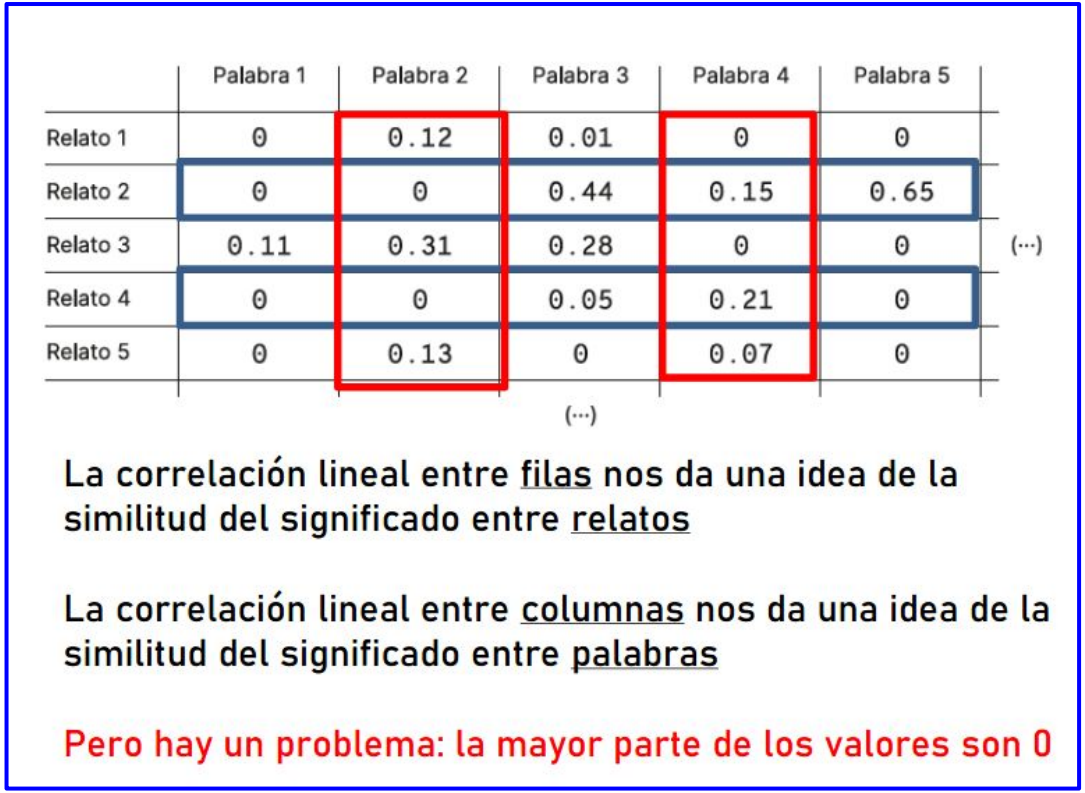
- Hasta aquí => conteo de palabras “crudos”, ponderados de alguna forma y/o mediante lexicones
- ¿Qué pasa si no queremos (o no podemos) usar lexicones? ¿Cómo detectamos los temas de un corpus sin leerlo y sin buscar palabras específicas?
- Tenemos un corpus documental muy grande y queremos una herramienta para hacer una primera “lectura” sin leer uno por uno los documentos.

¿Qué es?

- Las técnicas de modelado de tópicos apuntan eso: buscan detectar grupos o conjuntos de textos con una temática similar. Algoritmos basados en descomposición de matrices (NMF: Non-Negative Factorization) y modelos probabilísticos (LDA: Latent Dirichlet Allocation).

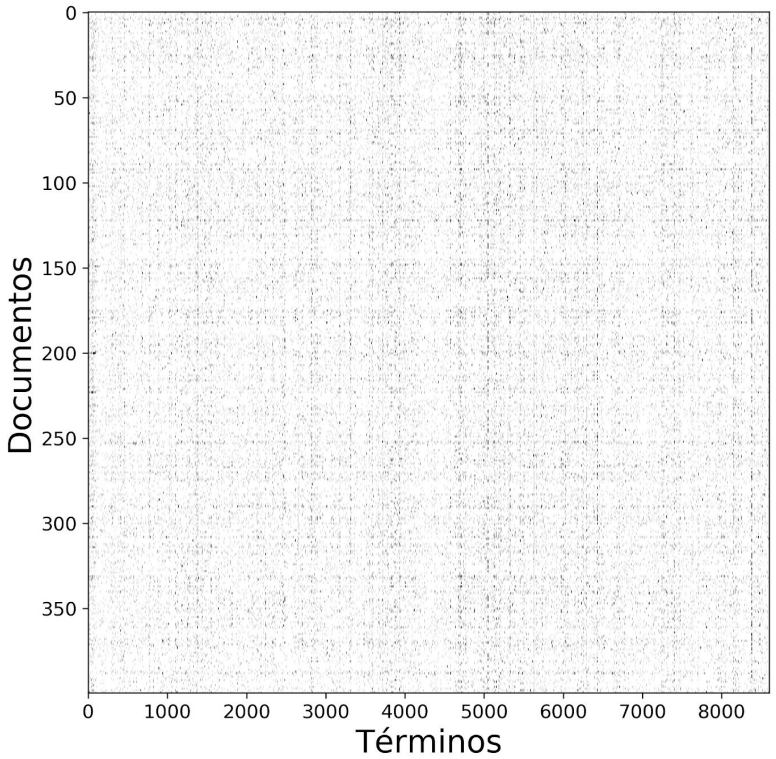
- Vamos a centrarnos en LDA.

- La matriz de documentos-términos suele tener muchos ceros
- Problema: se hace difícil medir la relación entre los distintos documentos o términos



	Palabra 1	Palabra 2	Palabra 3	Palabra 4	Palabra 5	
Relato 1	0	0.12	0.01	0	0	
Relato 2	0	0	0.44	0.15	0.65	
Relato 3	0.11	0.31	0.28	0	0	(···)
Relato 4	0	0	0.05	0.21	0	
Relato 5	0	0.13	0	0.07	0	
		(···)				

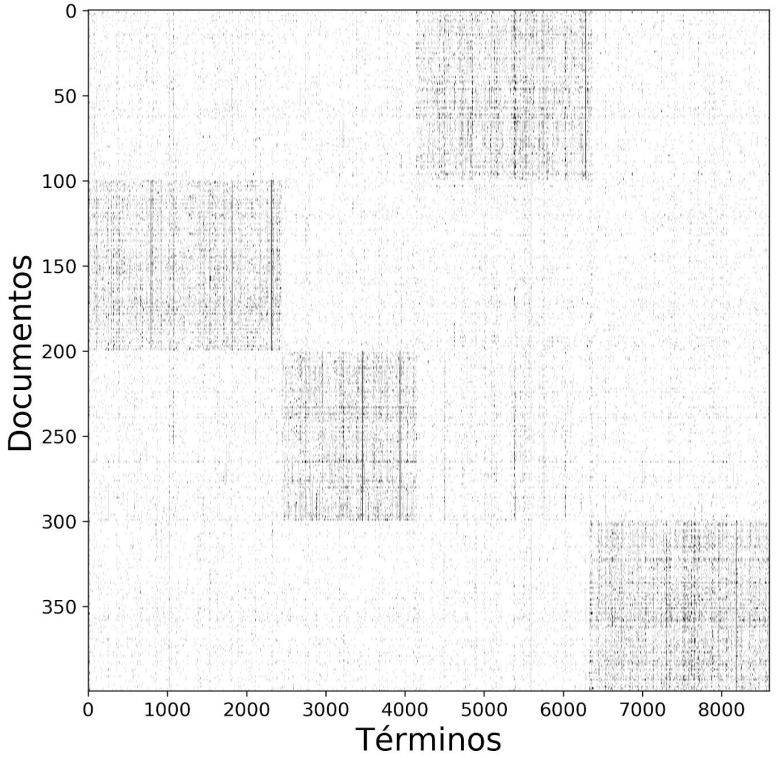
¿Qué es un tópico?



¿Cómo se ve una matriz de documentos por términos real?

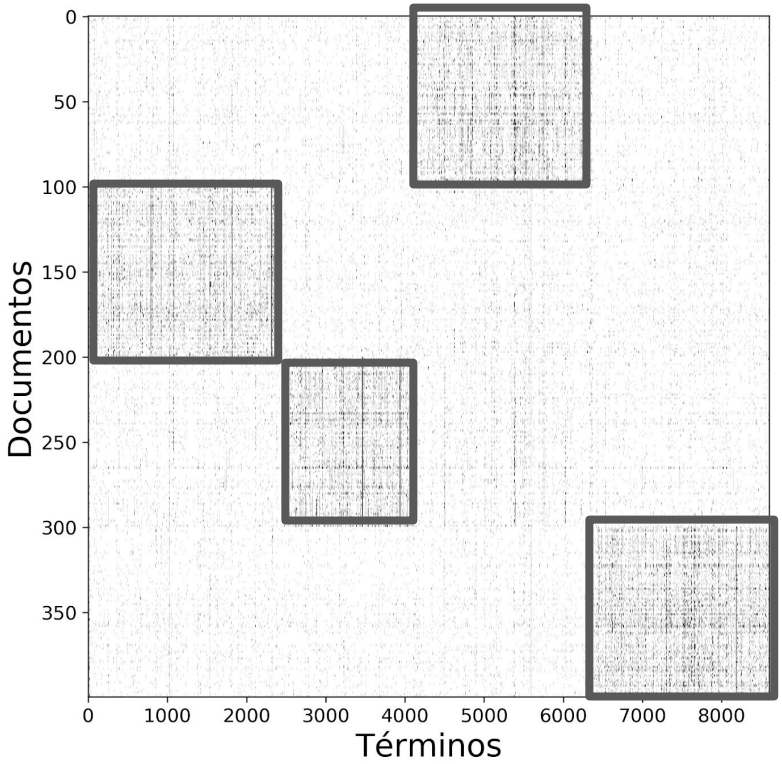
En blanco las componentes igual a cero; en negro las componentes distintas de cero.

¿Qué es un tópico?



Ordenando la matriz por filas y columnas…

¿Qué es un tópico?



Emergencia de bloques: Conjunto de **documentos que usan términos similares**. Estos bloques emergen naturalmente del "ordenamiento" de la matriz de documentos por términos.

A los bloques los identificamos como **tópicos** o **ejes temáticos**.

¿Cómo hacemos el ordenamiento? Algoritmos de detección de tópicos

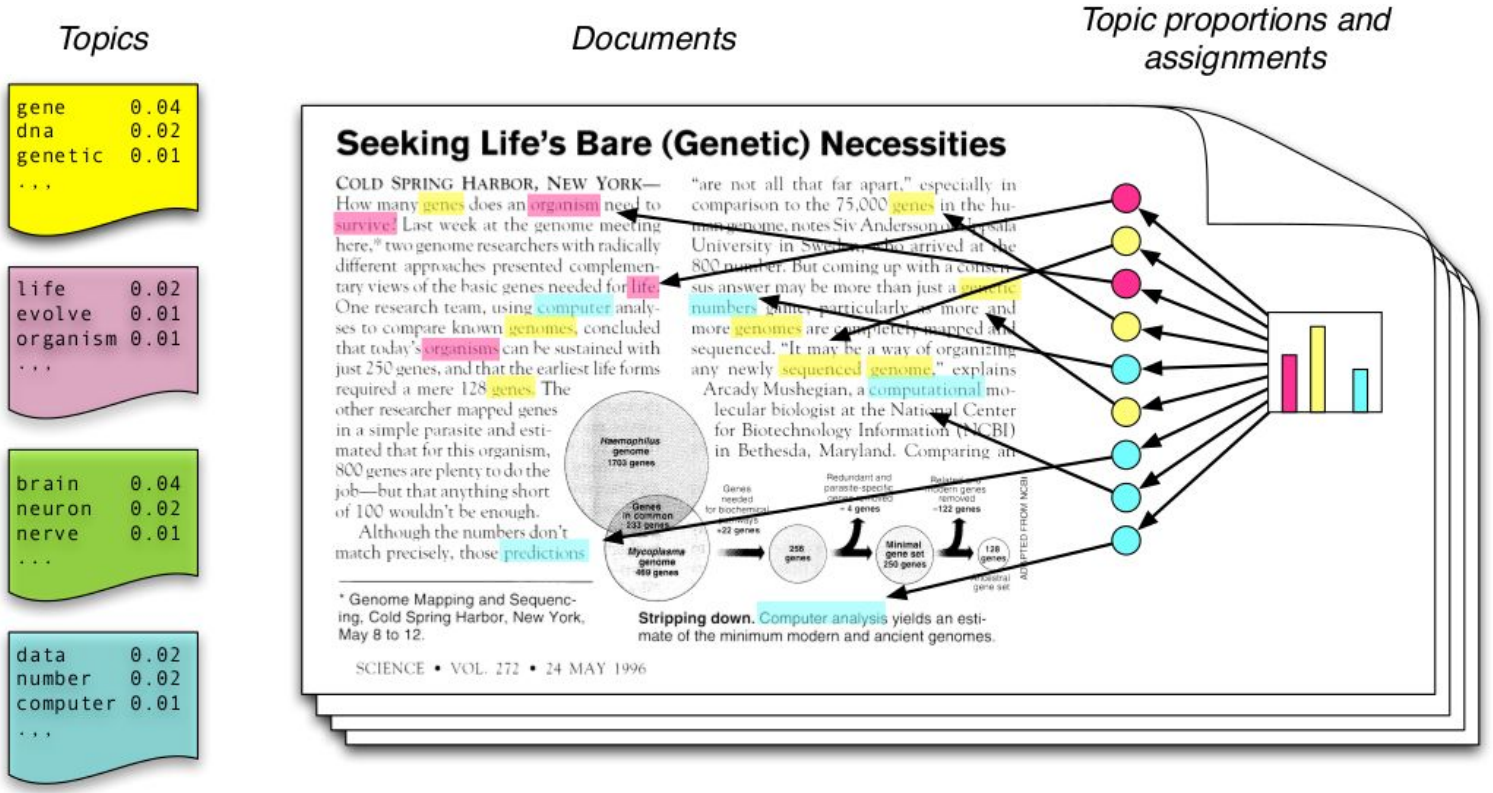
Latent Dirichlet Allocation (LDA)

Modelo probabilístico generativo (modelo para describir la forma en que se produjo la TFM)

Supuestos:

- un tópico es una distribución en el espacio de términos;
- un documento es una distribución en el espacio de tópicos (es una mixtura de tópicos).

Latent Dirichlet Allocation (LDA)

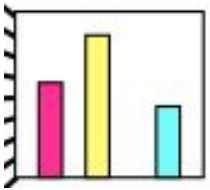


factor~data
IDAES_UNSAM

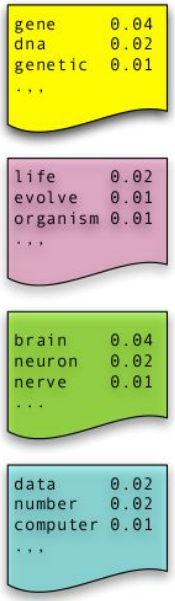
Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77-84.

Latent Dirichlet Allocation (LDA)

¿Cuál es el modelo generativo? La idea es ir construyendo término a término un documento. Supongamos que ya conocemos todas las distribuciones:



Elijo un tópico de la distribución del documento en el espacio de tópicos

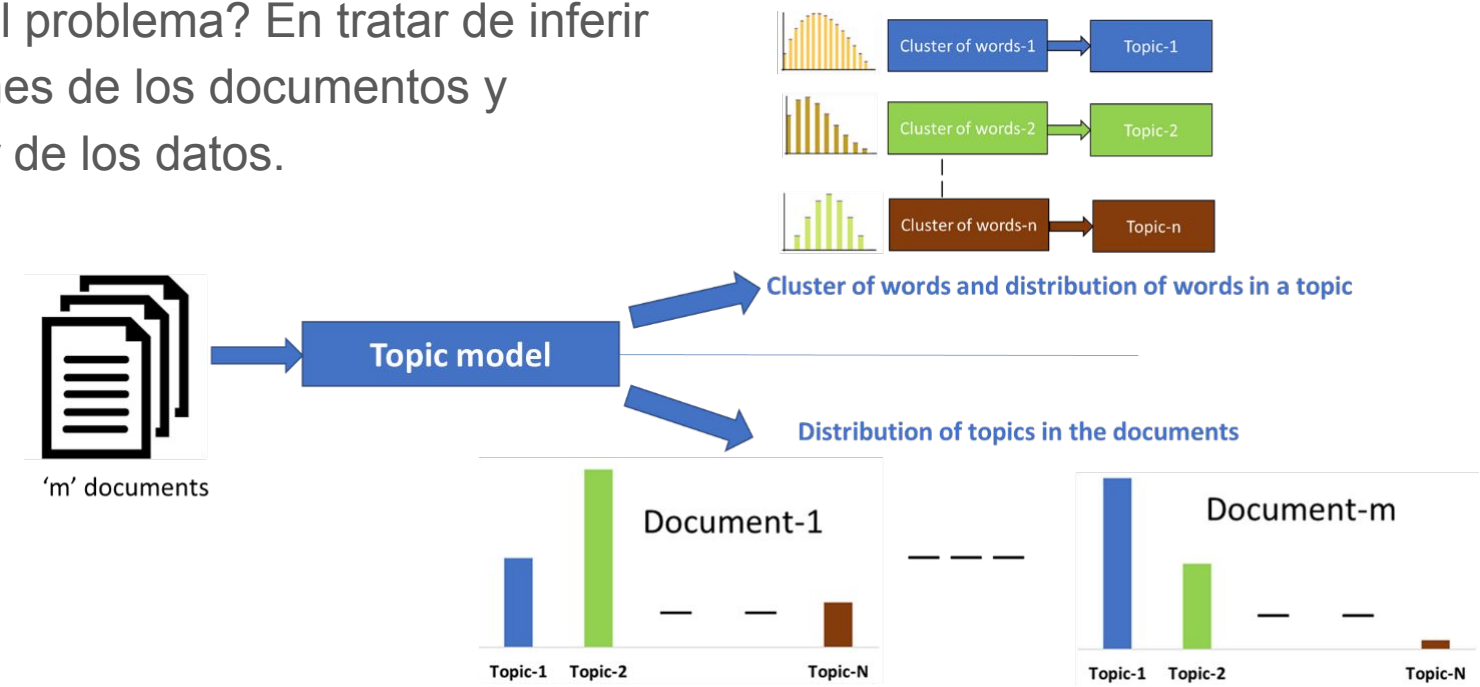


Elijo un término de la distribución del tópico elegido en el espacio de términos

El término elegido forma parte del documento e itero hasta completar los N términos del documentos

Latent Dirichlet Allocation (LDA)

¿Dónde está el problema? En tratar de inferir las distribuciones de los documentos y tópicos a partir de los datos.



factor~data
IDAES_UNSAM

https://www.mygreatlearning.com/blog/understanding-latent-dirichlet-allocation/

Latent Dirichlet Allocation (LDA)

Algoritmo

LDA assumes the following generative process for each document $\mathbf{w}$ in a corpus $\mathcal{D}$:

1. Choose $N \sim \text{Poisson}(\xi)$.
2. Choose $\theta \sim \text{Dir}(\alpha)$.
3. For each of the N words w_n:
   (a) Choose a topic $z_n \sim \text{Multinomial}(\theta)$.
   (b) Choose a word w_n from $p(w_n \mid z_n, \beta)$, a multinomial probability conditioned on the topic z_n.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, *3*, 993-1022.

Latent Dirichlet Allocation (LDA)

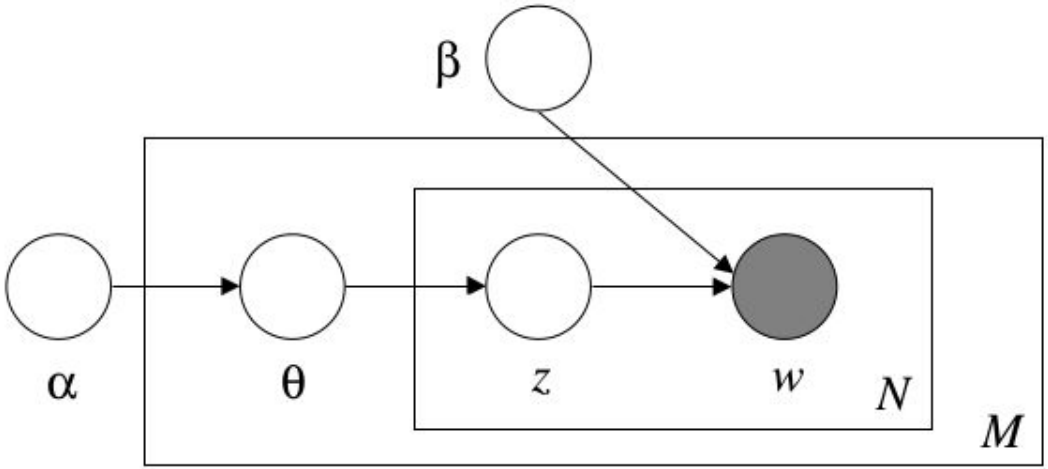


Figure 1: Graphical model representation of LDA. The boxes are "plates" representing replicates. The outer plate represents documents, while the inner plate represents the repeated choice of topics and words within a document.

Latent Dirichlet Allocation (LDA)

No vamos a ver en detalle la matemática ni los procesos de estimación pero esta ecuación da una intuición de lo que está pasando:

$$p(\mathbf{w}\,|\,\alpha,\beta) = \int p(\theta\,|\,\alpha)\left(\prod_{n=1}^{N}\sum_{z_n} p(z_n\,|\,\theta)\,p(w_n\,|\,z_n,\beta)\right) d\theta.$$

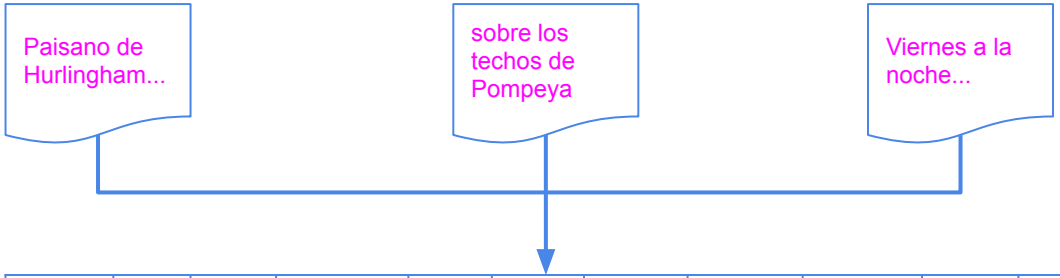
Proba de obtener un documento dado parámetros del modelo

Probabilidad de elegir el tópico del documento

Proba de elegir un término dado un tópico

Objetivo: inferir estos objetos (a través de inferir los parámetros de las distintas distribuciones).

Del texto crudo al texto como dato



	a	de	hurlingham	la	los	noche	paisano	pompeya	techos	vernes
C1	0	1	1	0	0	1	1	0	0	0
C2	0	1	0	0	1	0	0	1	1	0
C3	1	0	0	1	0	1	0	0	0	1

De la matriz de términos a la matriz de tópicos

	hurlingham	noche	paisano	pompeya	techos	vernes
C1	1	1	1	0	0	0
C2	0	0	0	1	1	0
C3	0	1	0	0	0	1

Matriz de Frecuencia de términos

Matriz de Documentos x Tópicos

	T1: Barrios	T2: Fiesta
C1	0.9	0.1
C2	0.9	0.1
C3	0.15	0.85

	hurlingham	noche	paisano	pompeya	techos	vernes
T1	0.8	0.4	0.8	0.9	0.6	0.2
T2	0.3	0.9	0.1	0.3	0.4	0.9

Matriz de Términos x Tópicos

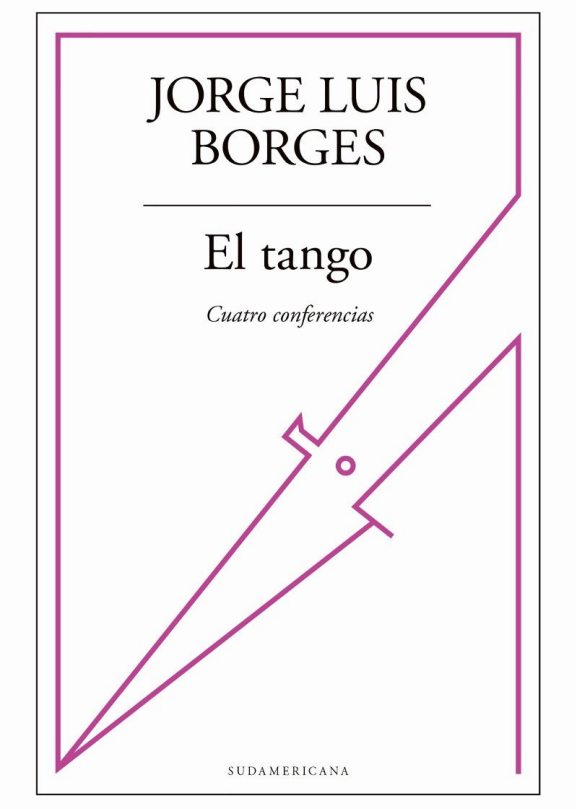
factor~data
IDAES_UNSAM

Latent Dirichlet Allocation (LDA)

Ventajas de los modelos generativos:

- Supuestos explícitas: si el modelo falla (por ejemplo, no encuentra los tópicos correctos en un corpus bien definido) se puede chequear si es porque los datos no cumplen alguna. De variar las hipótesis vienen las extensiones de LDA (STM vamos a ver la semana que viene).
- Generación de datos sintéticos y autoconsistencia: podemos inicializar el modelo con ciertos parámetros, generar datos sintéticos y ver si recuperamos los parámetros originales.

Un caso de aplicación (autobombo)



“El tango, como hemos visto, empezó, surge de la milonga, y es al principio un baile valeroso y feliz. Y luego, el tango va languideciendo y entristeciéndose…”

III Conferencia, p.80-81

Enfoque tradicional

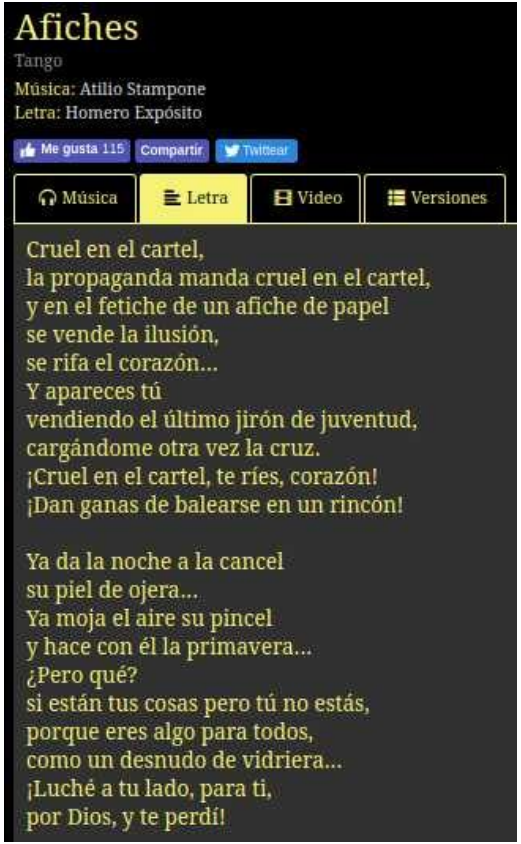
- Problema: analizar los temas de las letras de tango
- Enfoque “hermenéutico”: analizar pocas letras en profundidad
- Temas comunes: representaciones de género, figuras del “guapo”, representaciones del arrabal, etc.

Enfoque tradicional

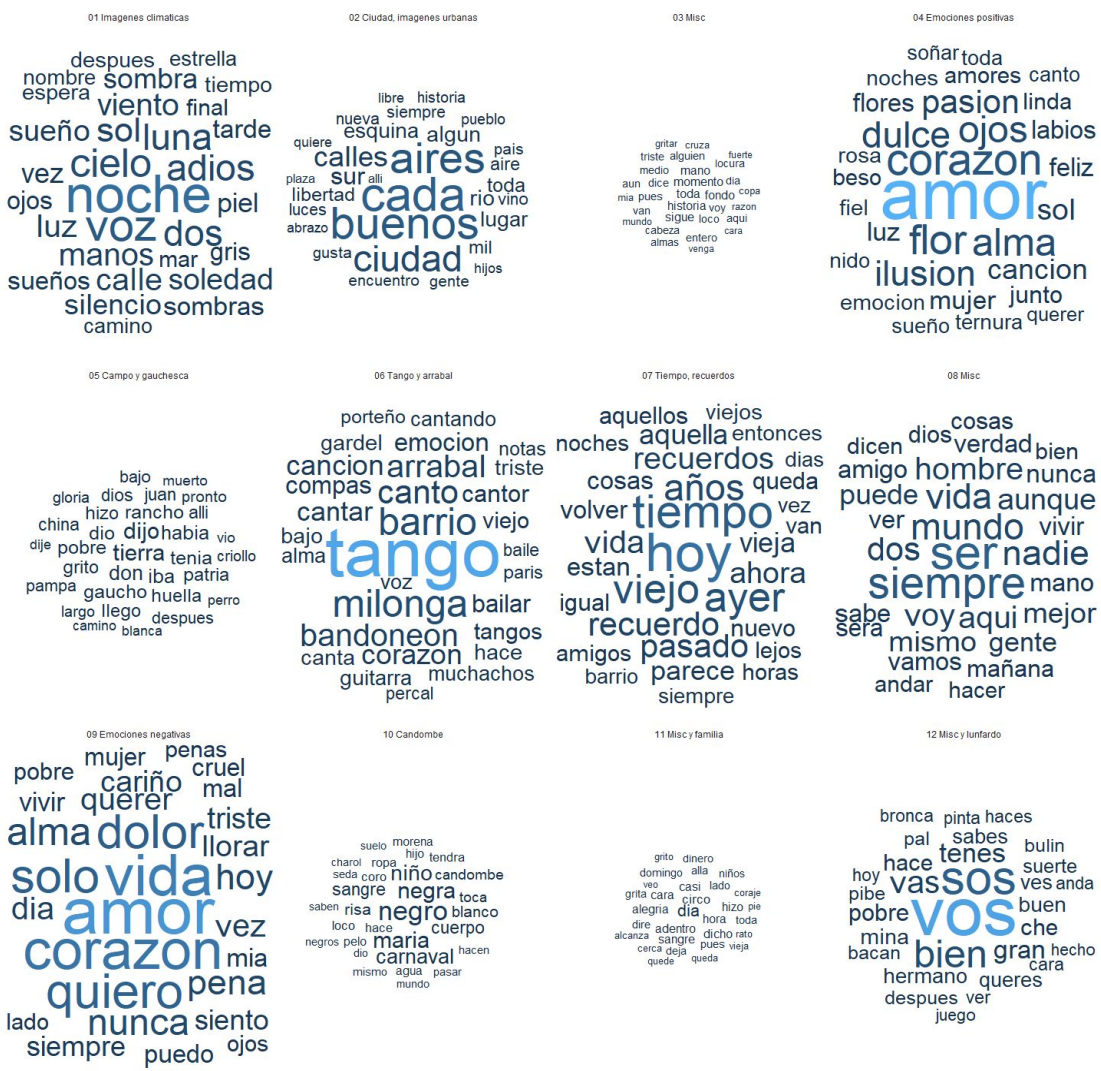
- Problema: analizar los temas de las letras de tango
- Enfoque “estadístico”
- Cantón (1972), analiza ciertos aspectos relevantes de las letras de los tangos cantados por Gardel

Enfoque propio

- Scrap de letras del sitio todotango.com
- Corpus: 5.700 letras
- Problema: analizar un corpus de ~5.700 letras de tango para detectar “tópicos” - Detección automática: Latent Dirichlet Allocation

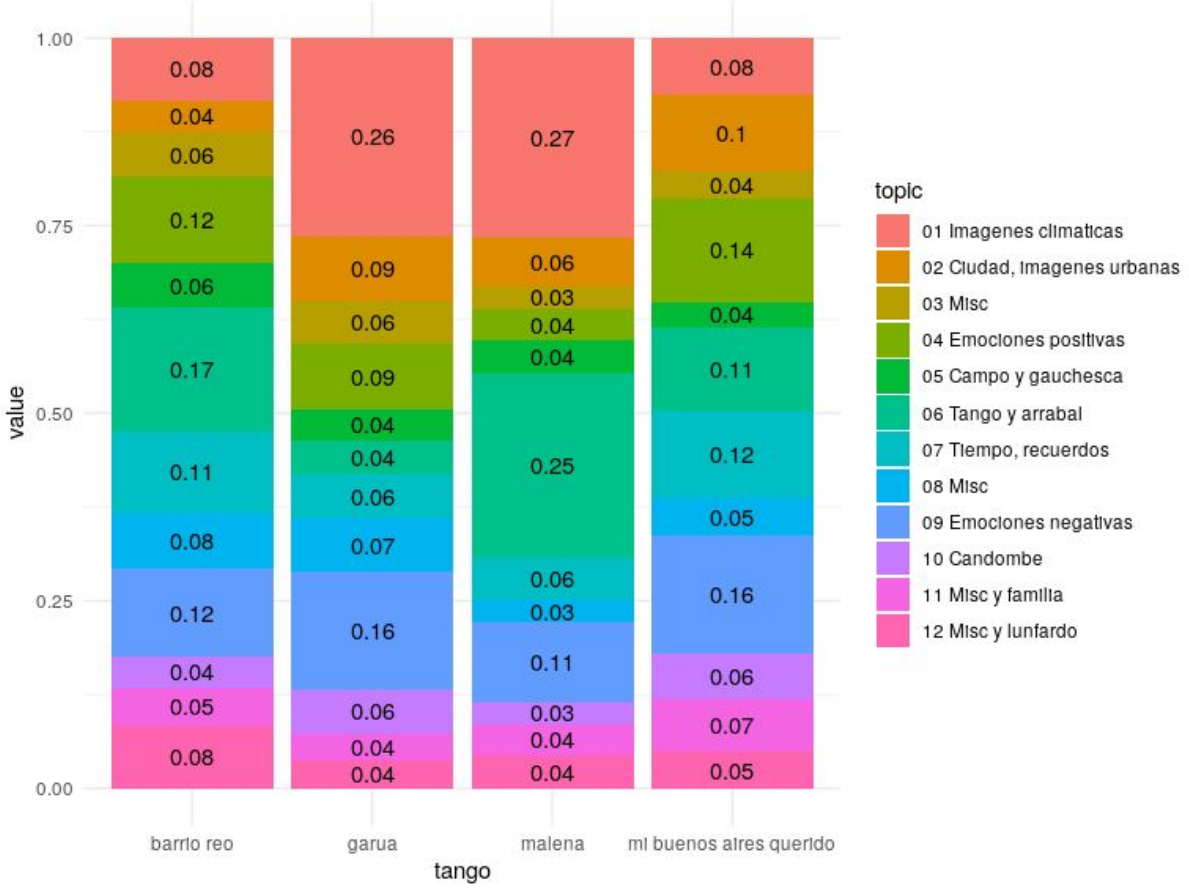


12 temas detectados

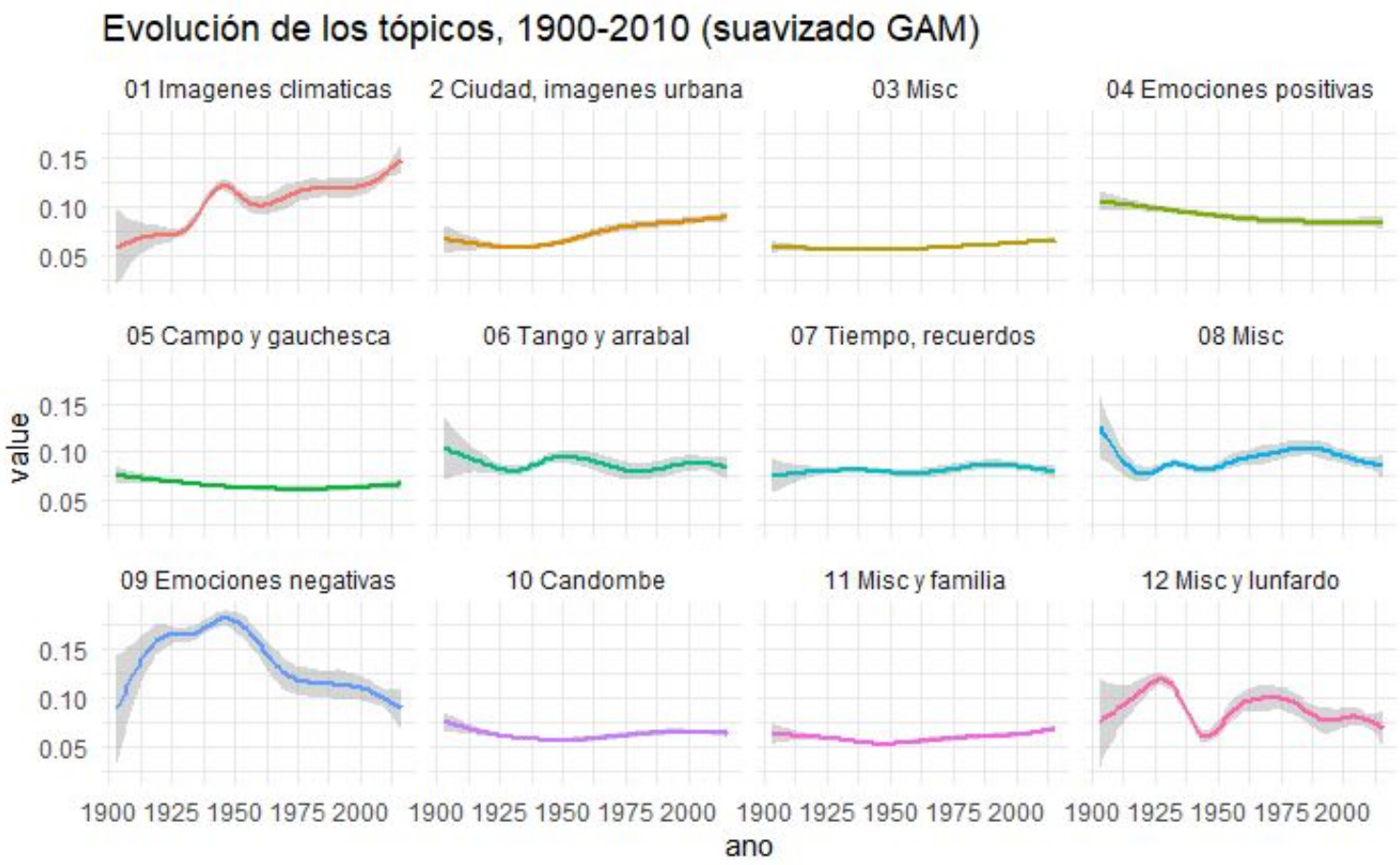


factor~data
IDAES_UNSAM

Composición de tópicos de 4 tangos



factor~data
IDAES_UNSAM



factor~data
IDAES_UNSAM

Resumen

- La TFM es un insumo para detectar tópicos
- Un “tópico” emerge como un de términos comunes usados por ciertos conjunto de documentos
- LDA es un método generativo para esa tarea

Vamos al notebook...