

# Introducción al Procesamiento de Lenguaje Natural

## Clase 3. Modelado de tópicos 1: Latent Dirichlet Allocation (LDA)



# ¿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, word) %>%  
  summarise(count=n())
```

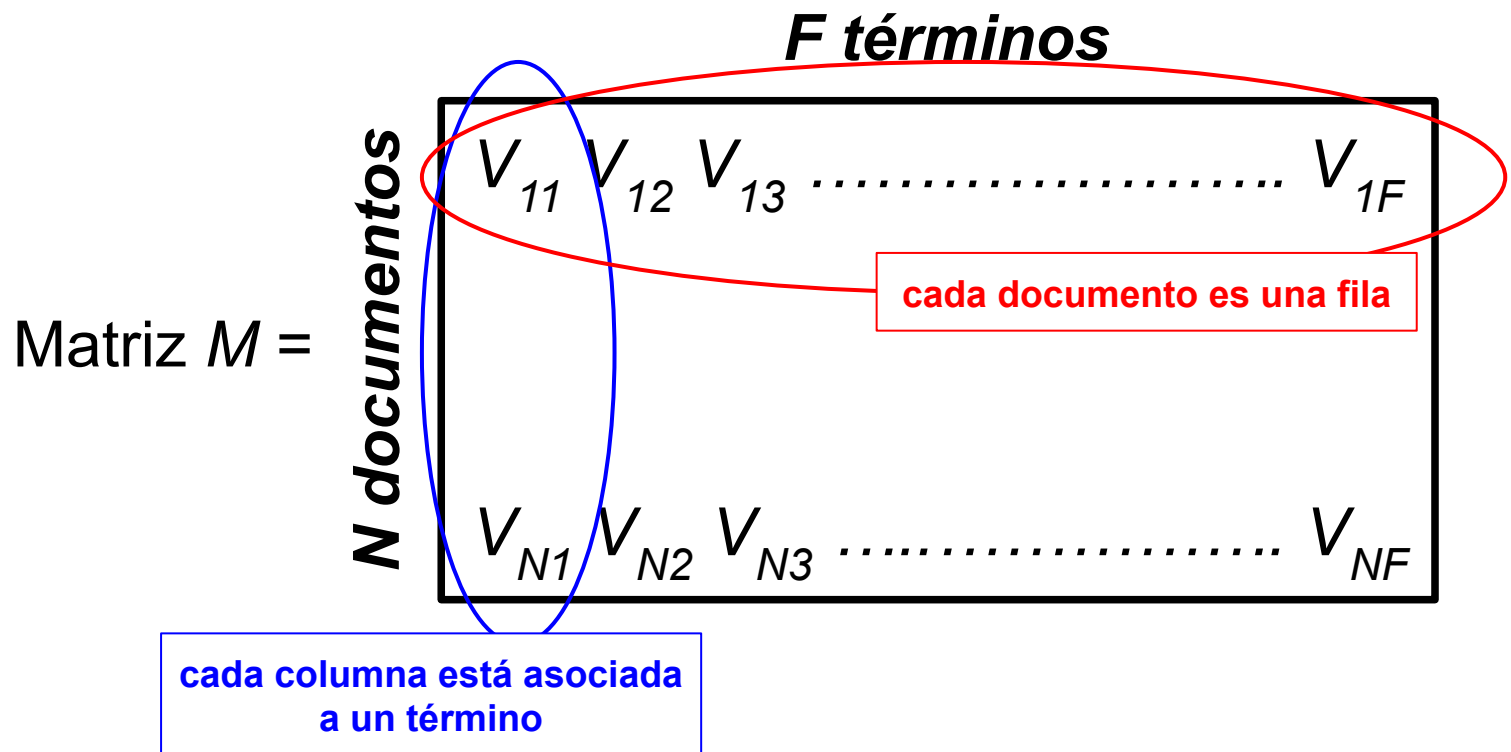
doc	word	count
1	no	1
1	es	1
1	la	2
1	conciencia	2
...	...	...
2	el	1
2	fantasma	2
2	del	1
2	comunismo	1



# 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)

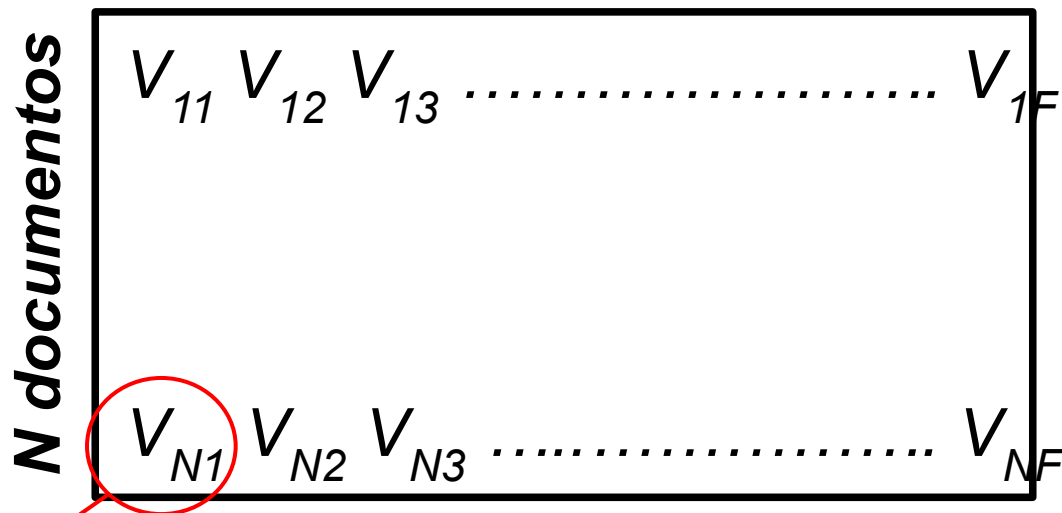


# Document-Term Matrix (TFM)

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

$F$  términos

Matriz  $M =$



Frecuencia del término

# 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 a 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).
  
- Hoy 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

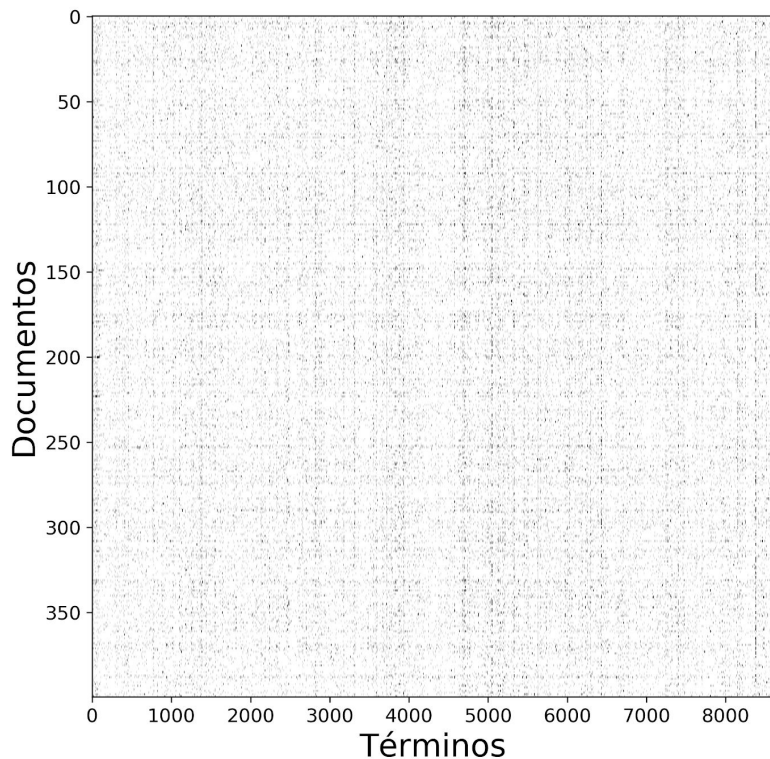
	Palabra 1	Palabra 2	Palabra 3	Palabra 4	Palabra 5	
Relato 1	0	0.12	0.01	0	0	
Relato 2	0	0	0.44	0.15	0.65	
Relato 3	0.11	0.31	0.28	0	0	(...)
Relato 4	0	0	0.05	0.21	0	
Relato 5	0	0.13	0	0.07	0	
			(...)			

La correlación lineal entre filas nos da una idea de la similitud del significado entre relatos

La correlación lineal entre columnas nos da una idea de la similitud del significado entre palabras

**Pero hay un problema: la mayor parte de los valores son 0**

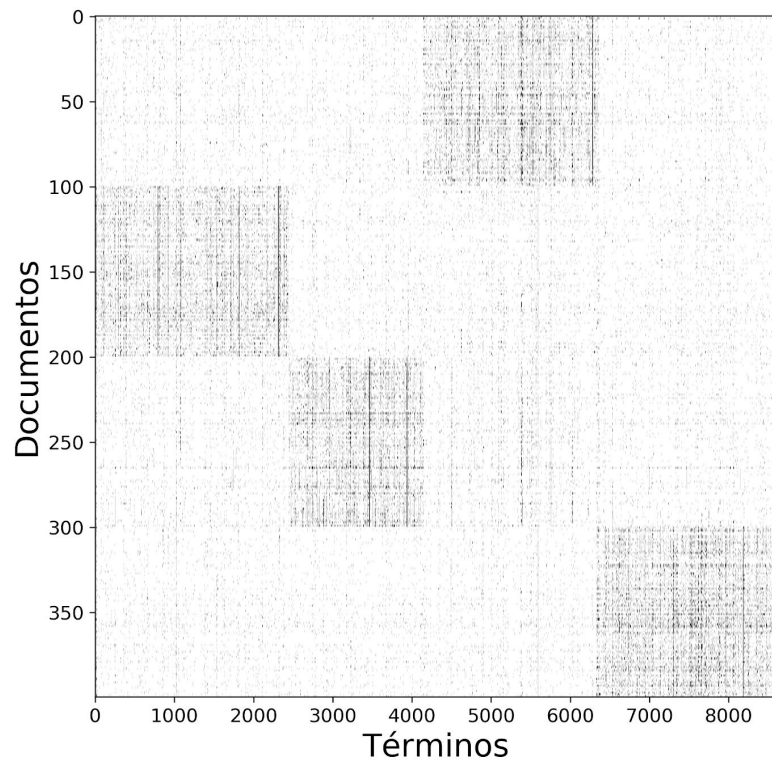
# ¿Qué es un tópico?



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

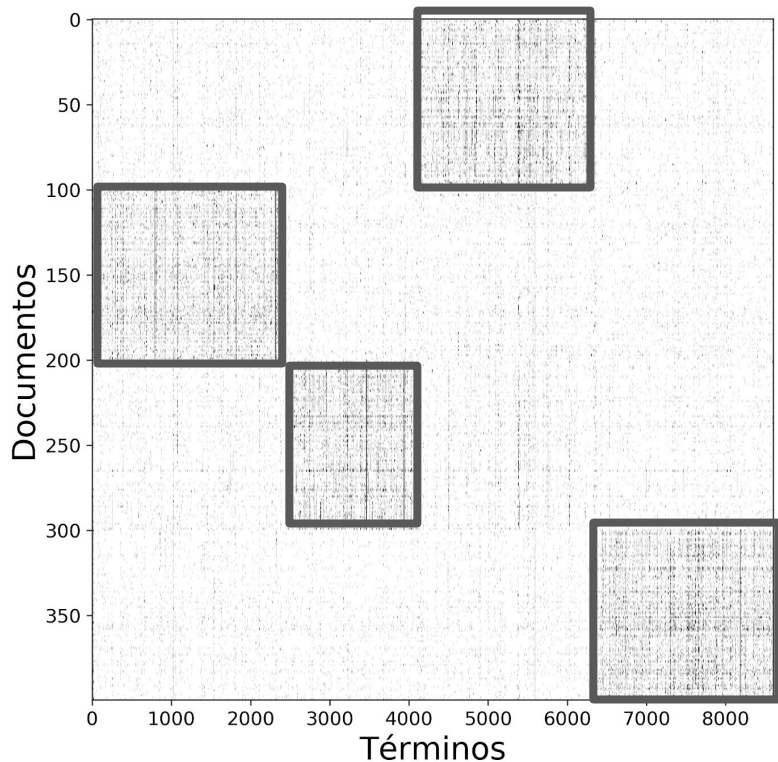
En blanco las componentes igual a cero; en negro las componentes distintas de cero.

# ¿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)

Topics

gene	0.04
dna	0.02
genetic	0.01
...	

life	0.02
evolve	0.01
organism	0.01
...	

brain	0.04
neuron	0.02
nerve	0.01
...	

data	0.02
number	0.02
computer	0.01
...	

Documents

## Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,\* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough. Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson at Uppsala University in Sweden. He also arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

**Stripping down.** Computer analysis yields an estimate of the minimum modern and ancient genomes.

\* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

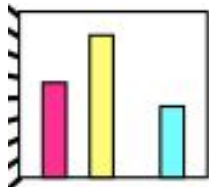
SCIENCE • VOL. 272 • 24 MAY 1996

ADAPTED FROM NCBI

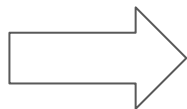
Topic proportions and assignments

# 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

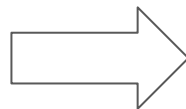


```
gene 0.04  
dna 0.02  
genetic 0.01  
...
```

```
life 0.02  
evolve 0.01  
organism 0.01  
...
```

```
brain 0.04  
neuron 0.02  
nerve 0.01  
...
```

```
data 0.02  
number 0.02  
computer 0.01  
...
```

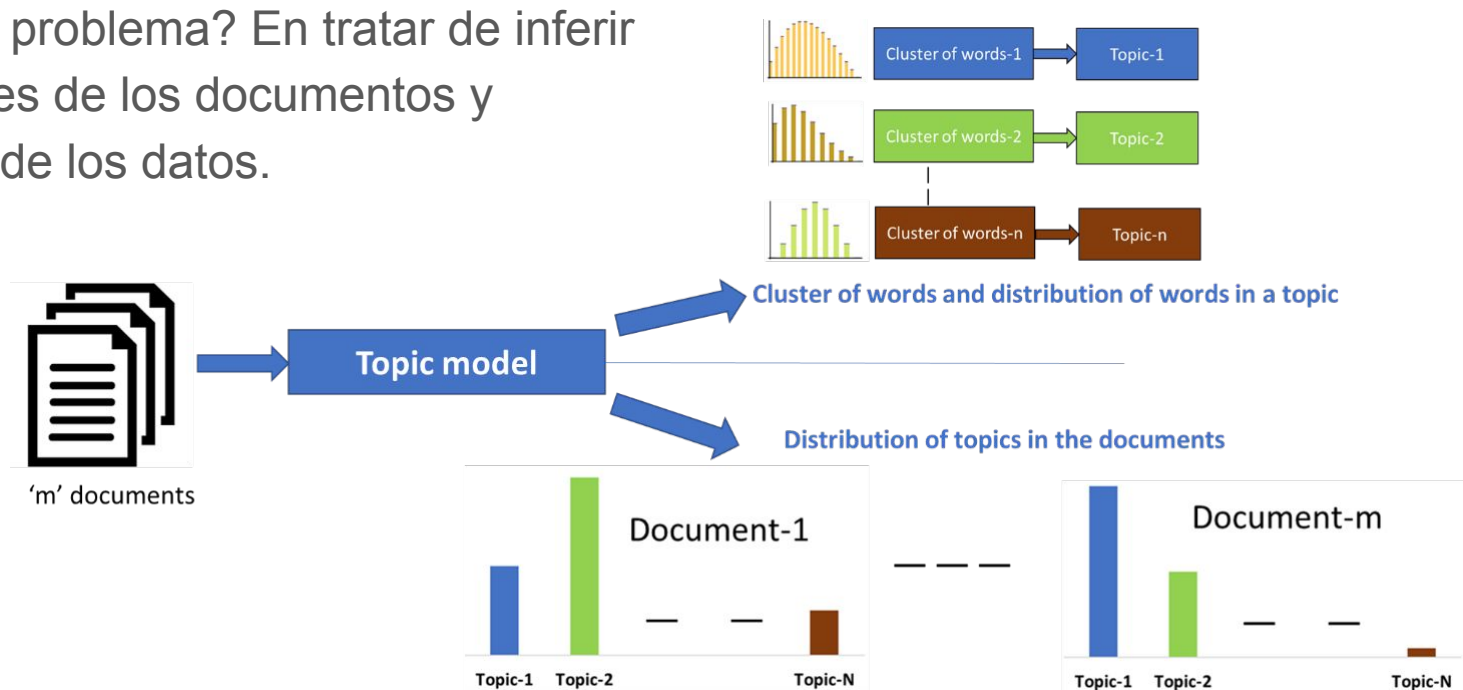


Elijo un término de la distribución del tópico elegido en el espacio de términos

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.



# Latent Dirichlet Allocation (LDA)

## Algorithm

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  $p(w_n | z_n, \beta)$ , 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)

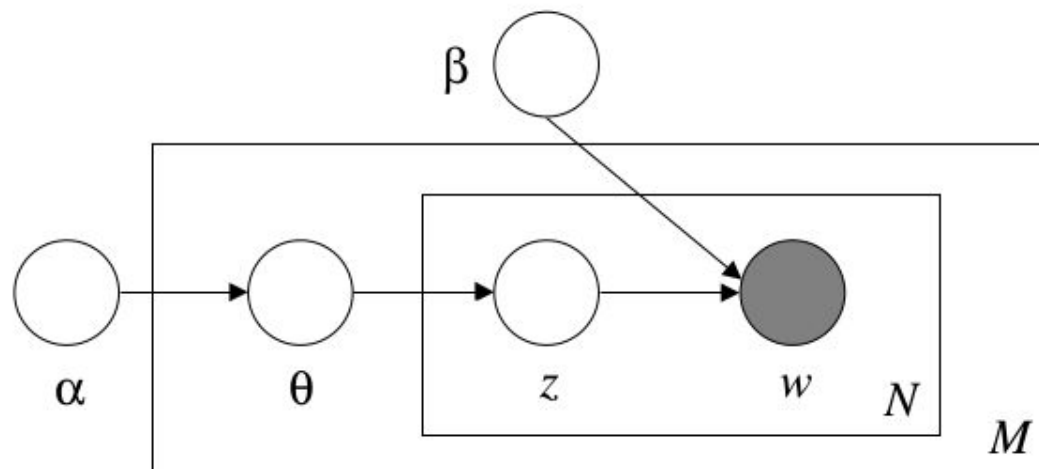


Figure 1: Graphical model representation of LDA. The boxes are “plates” representing replicates. The outer plate represents documents, while the inner plate represents the repeated choice of topics and words within a document.

# 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:

$$p(\theta, \mathbf{z}, \mathbf{w} | \alpha, \beta) = p(\theta | \alpha) \prod_{n=1}^N p(z_n | \theta) p(w_n | z_n, \beta),$$

Proba de obtener un documento dado parámetros del modelo

Probabilidad de la mezcla de tópicos del documento

Probabilidad de que ocurra ese tópico en la posición n del documento

Probabilidad de elegir un término n dado un tópico

# Latent Dirichlet Allocation (LDA)

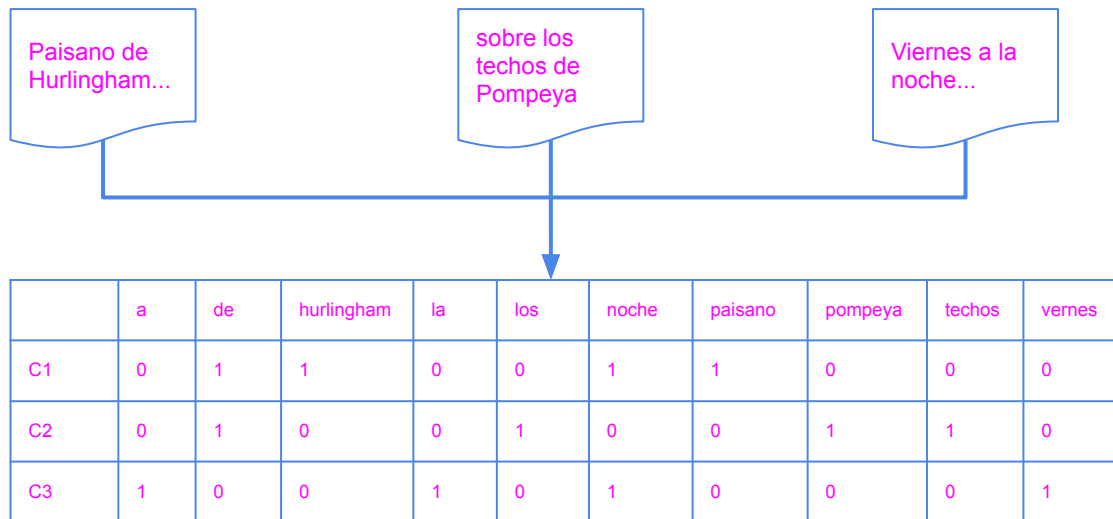
- Primero se define el perfil del documento, luego se asigna un t3pico a cada posici3n, y finalmente ese t3pico genera la palabra concreta.
- Dado un corpus de textos observados  $w$ , queremos inferir las distribuciones latentes:
  - distribuci3n de t3picos por documento ( $\theta$ )
  - asignaci3n de t3pico por palabra ( $z$ )
- Es un problema intratable.



# Latent Dirichlet Allocation (LDA)

- Métodos de estimación
  - Gibbs sampling
  - Variational Inference
  - Online LDA

# Del texto crudo al texto como dato



# 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

	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

Matriz de Documentos x Tópicos



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

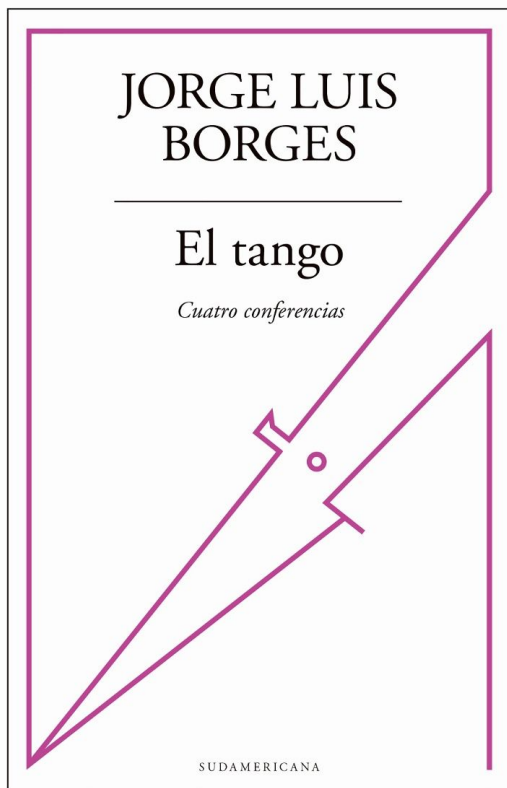


# Latent Dirichlet Allocation (LDA)

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.



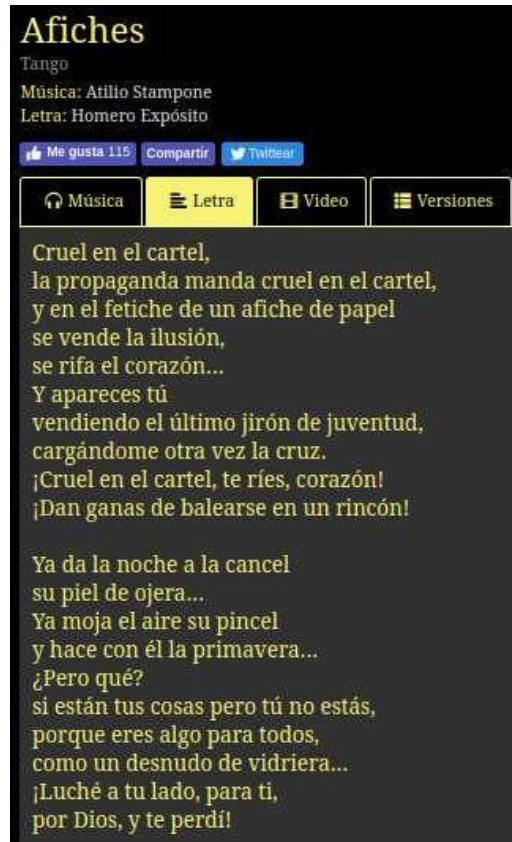
# Enfoque tradicional

- 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



A screenshot of a web page titled "Afiches" from the website todotango.com. The page displays the lyrics of a tango song. At the top, it identifies the genre as "Tango", the music by "Atilio Stampone", and the lyrics by "Homero Expósito". There are social media interaction buttons for "Me gusta 115", "Compartir", and "Twitter". Below these are navigation tabs for "Música", "Letra" (which is highlighted in yellow), "Video", and "Versiones". The lyrics are presented in a dark-themed interface with light-colored text.

**Afiches**  
Tango  
Música: Atilio Stampone  
Letra: Homero Expósito

Me gusta 115   Compartir   Twitter

Música   **Letra**   Video   Versiones

Cruel en el cartel,  
la propaganda manda cruel en el cartel,  
y en el fetiche de un afiche de papel  
se vende la ilusión,  
se rifa el corazón...  
Y apareces tú  
vendiendo el último jirón de juventud,  
cargándome otra vez la cruz.  
¡Cruel en el cartel, te ríes, corazón!  
¡Dan ganas de balearse en un rincón!

Ya da la noche a la cancel  
su piel de ojera...  
Ya moja el aire su pincel  
y hace con él la primavera...  
¿Pero qué?  
si están tus cosas pero tú no estás,  
porque eres algo para todos,  
como un desnudo de vidriera...  
¡Luché a tu lado, para ti,  
por Dios, y te perdí!



# 12 temas detectados



factor-data  
EIDAES\_UNSAM

01 Imágenes climáticas

despues estrella  
nombre sombra tiempo  
espera viento final  
sueño sol luna tarde  
vez cielo adios  
ojos noche piel  
luz voz dos  
manos mar gris  
sueños calle soledad  
silencios sombras  
camino

05 Campo y gauchesca

bajo muerto  
gloria dios juan pronto  
hizo rancho allí  
china dio dijo habia vio  
dije pobre tierra tenia criollo  
grito don iba patria  
pampa gaucho huella perro  
largo llevo despues  
camino blanca

09 Emociones negativas

mujer penas  
pobre carino cruel  
vivir querer triste  
alma dolor llorar  
solo vida hoy  
dia amor vez  
corazon mia  
quiero pena  
lado nunca siento  
siempre puedo ojos

02 Ciudad, imágenes urbanas

libre historia  
nueva siempre pueblo  
esquina algun  
quiere calles aires pais aire  
plaza sur ali toda  
libertad cada rio vino  
luzes abrazo buenos lugar  
gusta ciudad mil  
encuentro gente hijos

06 Tango y arrabal

porteño cantando  
gardel emocion notas  
cancion arrabal triste  
compas canto cantor  
cantar barrio viejo  
bajo alma tango baile paris  
voz milonga bailar  
bandoneon tangos  
canta corazon hace  
guitarra muchachos  
percal

10 Candombe

sueño morena  
charol ropa hijo tendra  
seda cora niño candombe  
sangre negra loca  
saben risa negro blanco  
loco hace cuerpo  
negros pelo maria hacen  
dio carnaval mundo  
mismo agua pasar mundo

03 Misc

gitar cruz fuerte  
triste alguien bcura  
medio mano  
aun dice momento dia  
mia pues toda fondo copa  
van historia voy razon  
mundo sigue loco aqui  
cabeza cara  
almas entero vengá

07 Tiempo, recuerdos

aquellos viejos  
noches aquella entonces  
recuerdos dias  
cosas años queda  
volver tiempo vez van  
vida hoy vieja  
están ahora  
igual viejo ayer  
recuerdo nuevo  
amigos pasado lejos  
barrio parece horas  
siempre

11 Misc y familia

grito dinero  
domingo alla niños  
veo casi lado coraje  
grita cara circo  
alegría dia hizo pie  
hora toda  
dise dentro dicho rato  
alcanza sangre pues vieja  
cerca deja queda queda

04 Emociones positivas

soñar toda  
noches amores canto  
flores pasión linda  
dulce ojos labios  
rosa corazon feliz  
beso amor sol  
fiel luz flor alma  
nido ilusión canción  
emocion mujer junto  
sueño ternura querer

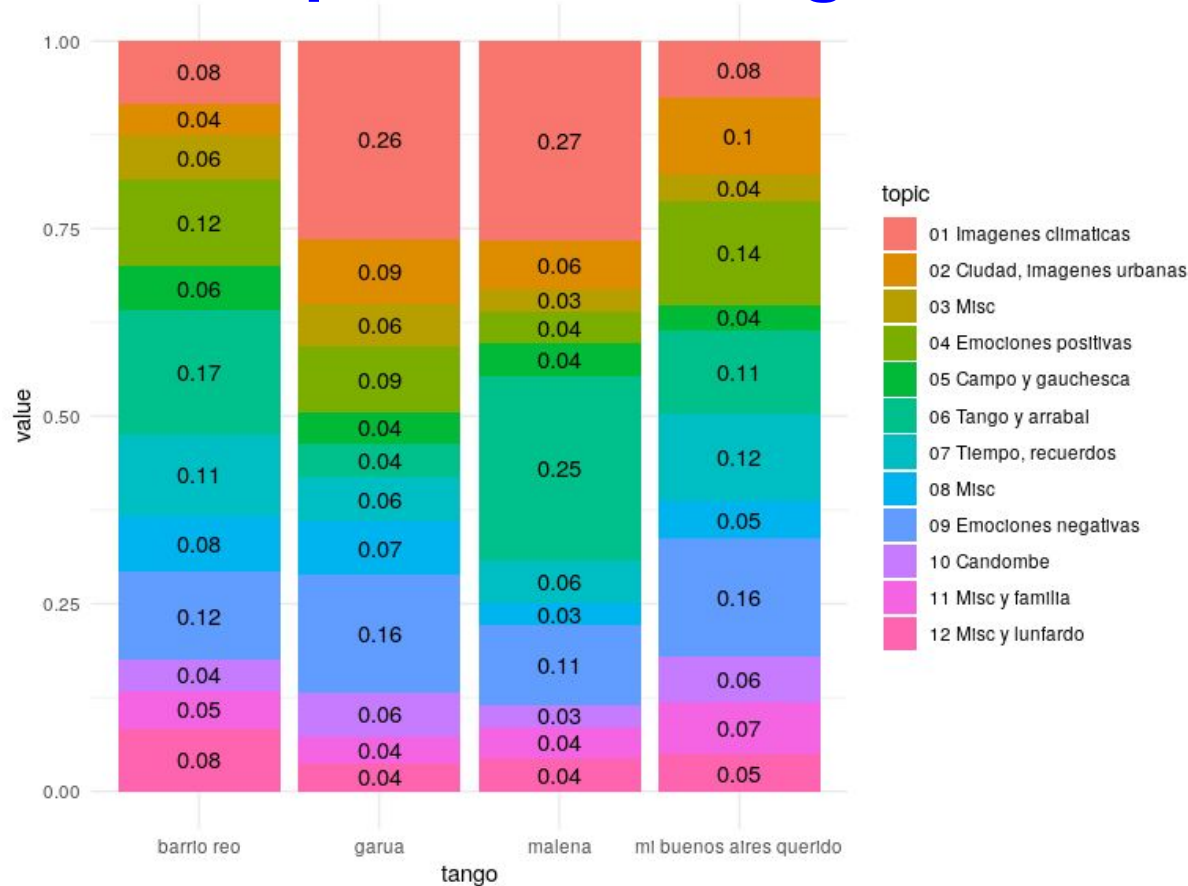
08 Misc

dicen dios cosas  
verdader bien  
amigo hombre nunca  
puede vida aunque  
ver mundo vivir  
dos ser nadie  
siempre mano  
sabe voy aqui mejor  
sera mismo gente  
vamos mañana  
andar hacer

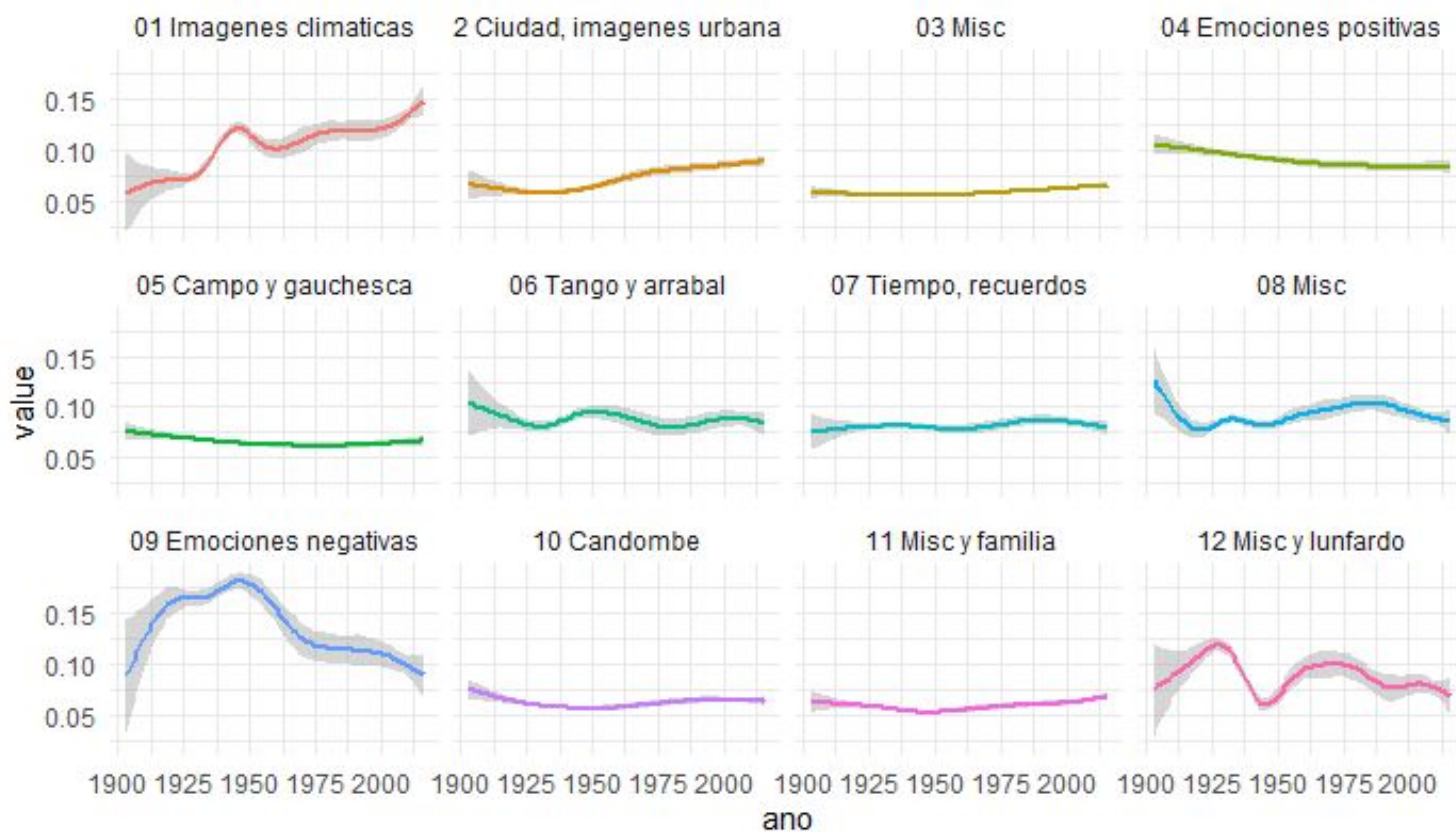
12 Misc y lunfardo

bronca pinta haces  
pal sabes bulin  
hace tenes suerte  
hoy vassos ves anda  
pibe vassos buen  
pobre vos che  
mina bien gran hecho  
bacan bien gran cara  
hermano queres  
despues ver  
juego

# Composición de tópicos de 4 tangos



## Evolución de los tópicos, 1900-2010 (suavizado GAM)



# Latent Dirichlet Allocation (LDA)

Ventajas de los modelos generativos:

- Supuestos explícitos: 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.
- Variando esas hipótesis/supuestos se producen las extensiones de LDA
- Structural Topic Modeling (STM), Hierarchical Topic Modeling, BERTopic (este lo vamos a ver más adelante)

# Vamos al notebook...



# Un flujo de trabajo posible para topic modeling

## 1. Definir objetivos

Adecuación entre problemas, preguntas y objetivos de investigación y supuestos del modelo (en este caso, STM/LDA u otros)

Definición conceptual de “tópico”

Ejemplo: “agenda setting”

# Un flujo de trabajo posible para topic modeling

1. Definir objetivos

2. Construir un corpus

- Corpus con una distribución “balanceada”
- Evitar sesgos (o al menos conocerlos y cuantificarlos)
- Curar los corpus
- Unidad de análisis (documentos, párrafos, oraciones, etc.)

# Un flujo de trabajo posible para topic modeling

1. Definir objetivos
2. Construir un corpus
3. Pre-procesar el texto
  - Filtrado de palabras
    - Stopwords por lista
    - Palabras de alta frecuencia
    - Eliminar caracteres
  - Reducción de complejidad (stemming/lemmatización, etc.)
  - Etc.

# Un flujo de trabajo posible para topic modeling

1. Definir objetivos
  2. Construir un corpus
  3. Pre-procesar el texto
  4. Seleccionar el modelo adecuado
- $K$  (nro. de tópicos) es el parámetro fundamental
  - Dispersión de tópicos y palabras son otros opcionales
  - Prueba de diferentes valores de  $K$
  - Algunas métricas (perplexity)

# Un flujo de trabajo posible para topic modeling

1. Definir objetivos
  2. Construir un corpus
  3. Pre-procesar el texto
  4. Seleccionar el modelo adecuado
  5. Interpretar y validar el modelo
- Analizar palabras por tópicos
    - Diferentes métricas (palabras más probables, FREX, etc.)
  - Muestrear documentos con altos valores en los diferentes tópicos y realizar una lectura cercana para validar
  - Colapsar tópicos similares (lógica similar a Grounded Theory)

# Un flujo de trabajo posible para topic modeling

1. Definir objetivos
  - Pocos estándares
  - Tabla de palabras por tópico
  - LDAviz
  - Diferentes métricas: relevancia, FREX, exclusividad, etc.
  - Tópicos genéricos vs tópicos exclusivos
2. Construir un corpus
3. Pre-procesar el texto
4. Seleccionar el modelo adecuado
5. Interpretar y validar el modelo
6. Comunicar el modelo.

# 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

