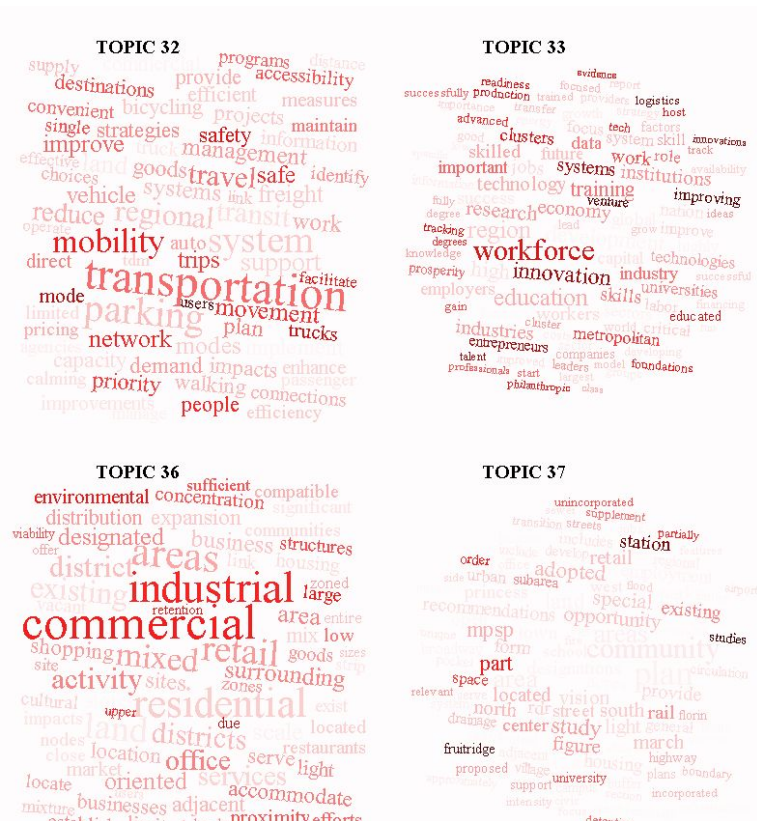


# **Topic Models**

Latent Dirichlet Allocation

# Tópic Modeling

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



# Idea...

En las grandes ciudades del mundo cada vez más personas, sobre todo jóvenes, eligen vivir en espacios más chicos pero funcionales, en edificios con servicios comunes, bien ubicados, conectados con el transporte público, cercanos a los lugares de trabajo.





...tuvieron que diseñar muebles funcionales a medida para aprovechar el espacio al máximo y deshacerse de cosas "superfluas"; también renunciar a la heladera con freezer, al horno y reemplazar la cama matrimonial por una de una plaza con carricama debajo,

# Idea...

En las grandes ciudades del mundo cada vez más personas, sobre todo jóvenes, eligen vivir en espacios más chicos pero funcionales, en edificios con servicios comunes, bien ubicados, conectados con el transporte público, cercanos a los lugares de trabajo.

...tuvieron que diseñar muebles funcionales a medida para aprovechar el espacio al máximo y deshacerse de cosas "superfluas"; también renunciar a la heladera con freezer, al horno y reemplazar la cama matrimonial por una de una plaza con carricama debajo,

## Viendo las palabras principales:

-  ¿Podemos **descubrir** de qué habla el documento?
-  ¿Hace falta leerlas en **orden** para determinar lo anterior?
-  ¿Habla de un sólo tema o hay **varios temas** a la vista?
-  ¿Algún tema **sobresale** más que otros?

# ¿Qué es un tópico?

Un tópico es **una distribución probabilística sobre el vocabulario**:  
Asignación de probabilidad a cada palabra.

Ejemplo, ¿qué tópicos serán los siguientes?

???????	
jugador	0.098
partido	0.095
pelota	0.094
...	
elección	0.001
jurado	0.001
...	

???????	
elección	0.098
voto	0.095
partido	0.090
...	
sandwich	0.040
jurado	0.005
pelota	0.001
...	

???????	
sandwich	0.095
jurado	0.094
horario	0.090
...	
pelota	0.010
voto	0.003
...	

# ¿Qué es un tópico?

Un tópico es **una distribución probabilística sobre el vocabulario**:  
Asignación de probabilidad a cada palabra.

Ejemplo, ¿qué tópicos serán los siguientes?

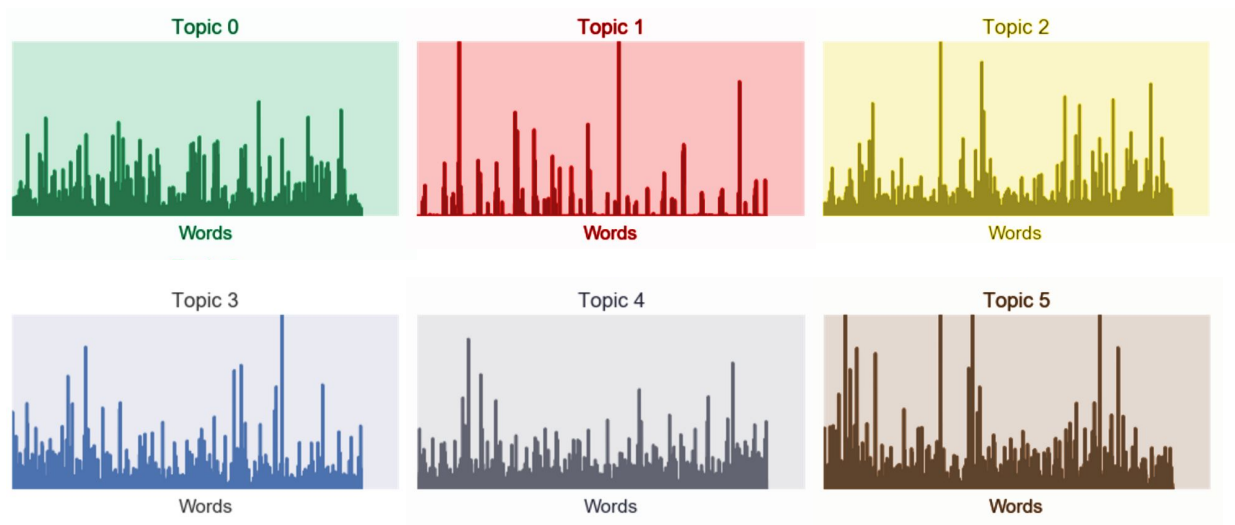
“Deportes”	
jugador	0.098
partido	0.095
pelota	0.094
...	
elección	0.001
jurado	0.001
...	

“Política”	
elección	0.098
voto	0.095
partido	0.090
...	
sandwich	0.040
jurado	0.005
pelota	0.001
...	

“Defensa tesis”	
sandwich	0.095
jurado	0.094
horario	0.090
...	
pelota	0.010
voto	0.003
...	

# ¿Qué es un tópico?

Un tópico es **una distribución probabilística sobre el vocabulario**:  
Asignación de probabilidad a cada palabra.



# Latent Dirichlet Allocation (LDA)

David Blei, Andrew Ng, Michael I. Jordan, 2003

Modelo probabilístico basada en las ideas de un proceso generativo.



# Latent Dirichlet Allocation (LDA)

David Blei, Andrew Ng, Michael I. Jordan, 2003

Modelo probabilístico basada en las ideas de un proceso generativo.

Topics

gene 0.04  
dna 0.02  
genetic 0.01  
...

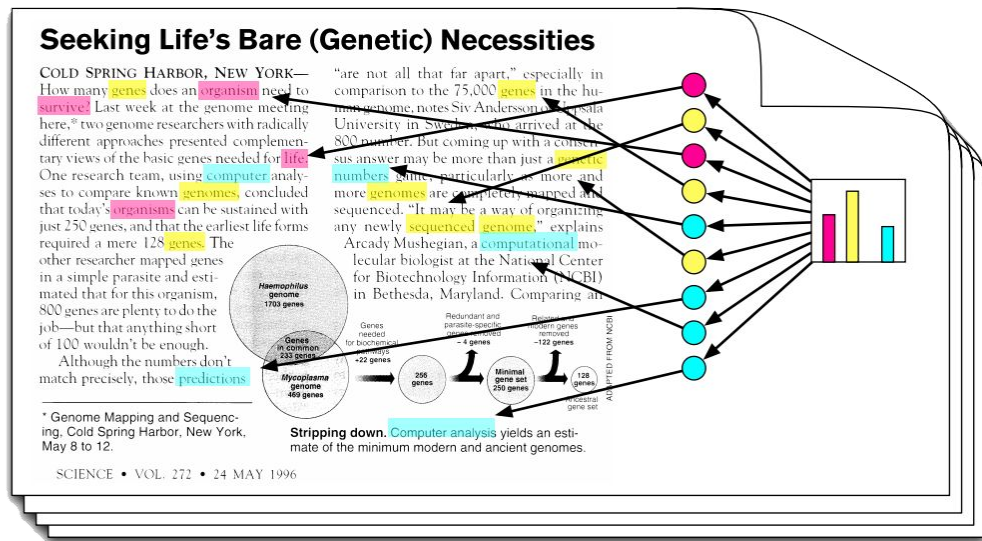
life 0.02  
evolve 0.01  
organism 0.01  
...

brain 0.04  
neuron 0.02  
nerve 0.01  
...

data 0.02  
number 0.02  
computer 0.01  
...

Documents

Topic proportions & assignments



# Latent Dirichlet Allocation (LDA)

David Blei, Andrew Ng, Michael I. Jordan, 2003

Modelo probabilístico basada en las ideas de un proceso generativo.

## Algoritmo Generativo

Idea: Generar documentos.

1. Creamos un documento vacío.
2. Asignamos al documento, una **proporción** de los distintos tópicos que tenemos.
3. Generamos palabras siguiendo el procedimiento:
  - **Elegimos un tópico** al azar (teniendo en cuenta las proporciones del ítem 2.).
  - Dado el tópico, **elegimos una palabra** según las probabilidades en ese tópico.
  - Repetimos desde (3) hasta generar **N palabras**
4. Repetimos desde el paso (1) hasta generar **D documentos**.

Un poco de deportes, un poco más de ciencia, un poco de política, etc.

Por ej, si el tópico era deportes, la palabra puede ser NBA.

# Pero....

Toda la estructura de la que hablamos **no está visible**. Hay que encontrarla.

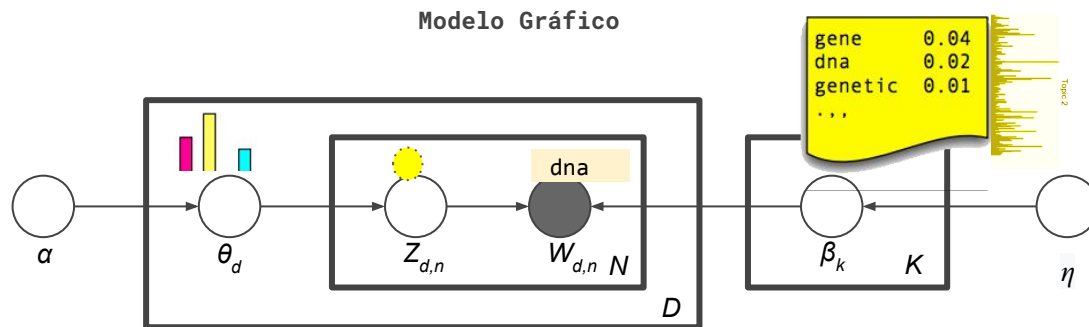
**Problema:**

Inferir los parámetros de un modelo probabilístico que explique lo mejor posible nuestros documentos.

**Luego:**

Podremos utilizar esa información para anotar y visualizar nuestros datos.

# Modelo LDA



Donde:

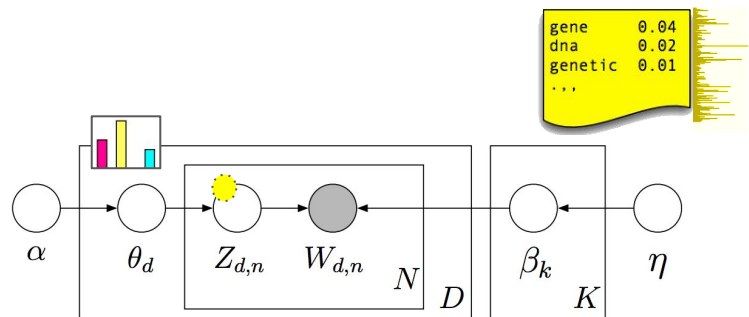
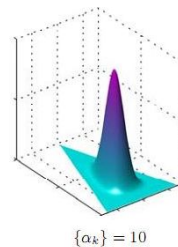
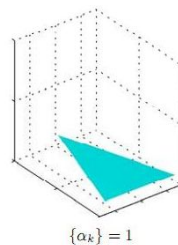
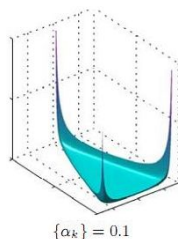
- $\alpha$  y  $\eta$  son hiperparámetros del modelo (dan más o menos flexibilidad a las proporciones de tópicos y a los tópicos)
- $\theta_d$  es la proporción de tópicos asignadas al documento  $d$
- $Z_{d,n}$  es el tópico asignado a la  $n$ -ésima palabra
- Cada  $\beta_k$  (tópico) es una distribución sobre las palabras del vocabulario
- $W_{d,n}$  representa la  $n$ -ésima palabra del documento  $d$

# Más preciso

## Modelo generativo

1. For each set  $d_m, m \in \{1 \dots M\}$ , choose a  $K$ -dimensional latent variable weight vector  $\theta_m$  from the Dirichlet distribution with scaling parameter  $\alpha$ :  $p(\theta_m | \alpha) = \text{Dir}(\alpha)$
2. For each discrete item  $w_n, n \in \{1 \dots N\}$  in set  $d_m$ 
  - (a) Draw a latent variable  $z_n \in \{1 \dots K\}$  from the multinomial distribution  $p(z_n = k | \theta_m)$
  - (b) Given the latent variable, draw a symbol from  $p(w_n | z_n, \beta)$ , where  $\beta$  is a  $V \times K$  matrix and  $\beta_{ij} = p(w_n = i | z_n = j, \beta)$

$$\begin{aligned}
 Z_{d,n} | \theta_d &\sim \text{Multinomial}(\theta_d) \\
 W_{d,n} | Z_{d,n}, \beta &\sim \text{Multinomial}(\beta_{Z_{d,n}}) \\
 \beta_k | \eta &\sim \text{Dirichlet}(\eta) \\
 \theta_d | \alpha &\sim \text{Dirichlet}(\alpha)
 \end{aligned}$$



$\hat{p}(\theta, z, \beta | w)$ ?

# ¿Cómo inferir entonces?

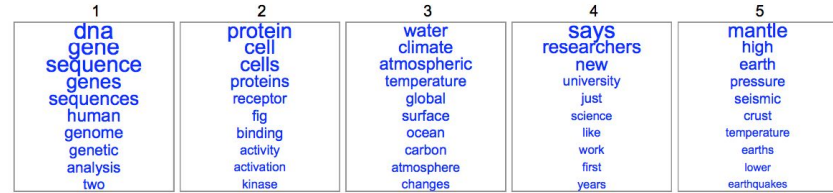
Algoritmos para aproximar la posterior:  $p(\theta, z, \beta \mid w)$ :

- Mean field variational methods (Blei et al., 2001, 2003)
- Expectation propagation (Minka and Lafferty, 2002)
- Collapsed Gibbs sampling (Griffiths and Steyvers, 2002)
- Distributed sampling (Newman et al., 2008; Ahmed et al., 2012)
- Collapsed variational inference (Teh et al., 2006)
- Online variational inference (Hoffman et al., 2010)
- Factorization based inference (Arora et al., 2012; Anandkumar et al., 2012)

# ¿Cómo se puede usar?

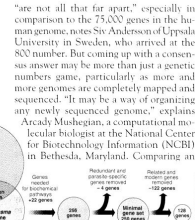
Luego de fitear los parámetros, podemos utilizar las distribuciones aproximadas para:

- (1) Visualizar las palabras más probables para cada tópico
- (2) Extraer la proporción de tópicos de cada documento



## Seeking Life's Bare (Genetic) Necessities

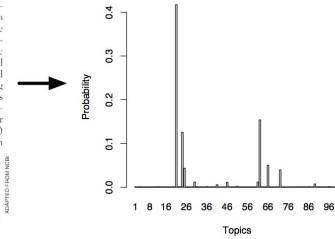
**COLD SPRING HARBOR, NEW YORK—**How many genes does an organism need to survive? Last week at the genome meeting here,\* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare 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



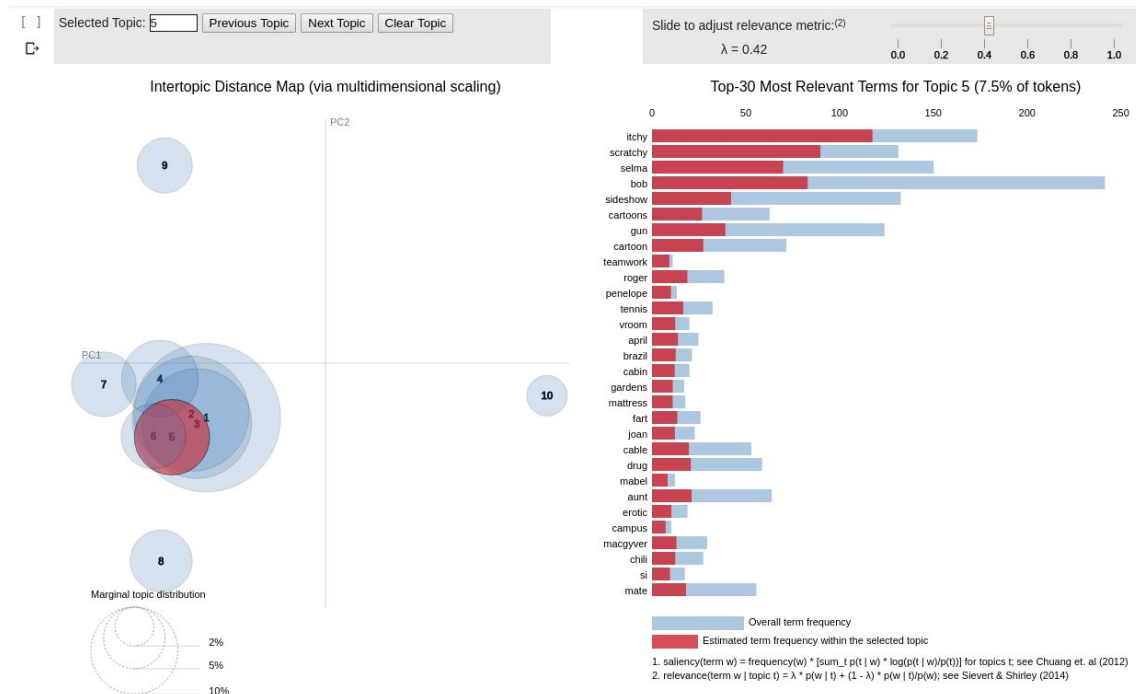
\* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 6 to 12.

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

SCIENCE • VOL. 272 • 24 MAY 1996



# Visualización de tópicos: LDAvis





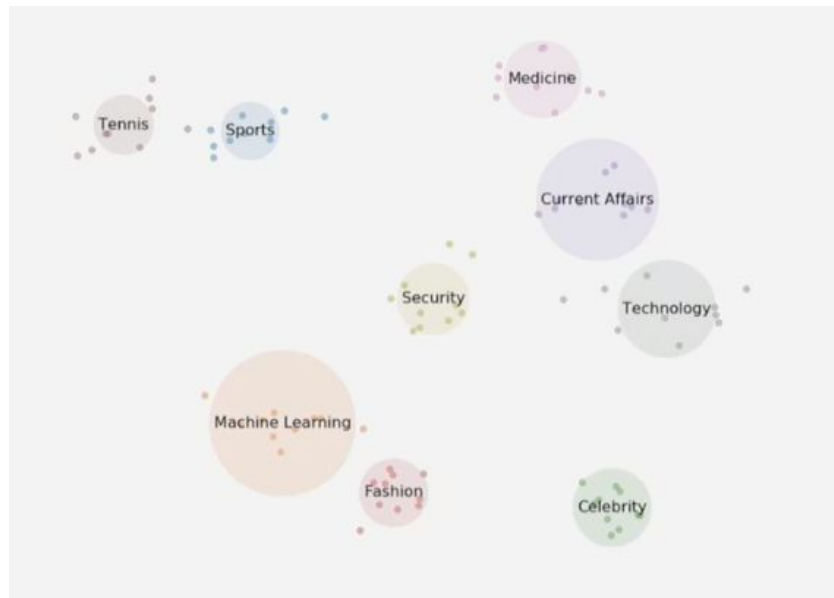
# ¿Aprendió bien?

## Necesitamos saber si...

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

## Métodos

- Métodos a ojimetro.
- Métodos intrínsecos.
- Métricas basadas en Human Judgments.
- Métodos extrínsecos (evaluar sobre una tarea)



# Método a ojometro

- Me fijo si las  $k$  palabras de cada tópico representan efectivamente un tópico de interés. En caso afirmativo le pongo un “label” que lo identifique.
- Uso los tópicos que me interesan para mi análisis

“Deportes”	
jugador	0.098
partido	0.095
pelota	0.094
...	
elección	0.001
jurado	0.001
...	

“Política”	
elección	0.098
voto	0.095
partido	0.090
...	
sandwich	0.040
jurado	0.005
pelota	0.001
...	

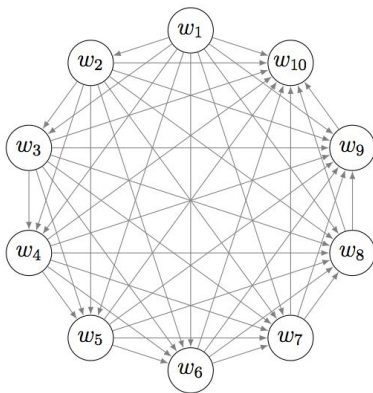
“Defensa tesis”	
sandwich	0.095
jurado	0.094
horario	0.090
...	
pelota	0.010
voto	0.003
...	

???????	
pomelo	0.103
lavarropa	0.094
barco	0.083
...	
caja	0.010
caramelo	0.003
...	

# Medidas Intrínsecas

## Topic coherence

- Tomo los top- $n$  palabras del tópico  $k$



Calculado con un corpus externo

$$C_{\text{UCI}} = \frac{2}{N \cdot (N - 1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^N \text{PMI}(w_i, w_j)$$

Topico  $k=\{\textit{game, sport, ball, team}\}$

$$C_{\text{UCI}} = \frac{1}{6} \cdot (\text{PMI}(\textit{game, sport}) + \text{PMI}(\textit{game, ball}) \\ + \text{PMI}(\textit{game, team}) + \text{PMI}(\textit{sport, ball}) \\ + \text{PMI}(\textit{sport, team}) + \text{PMI}(\textit{ball, team}))$$

# Métricas basadas en Human Judgments

## Word intrusion

- Dado un tópico, tomar las 5 palabras más probables
- Agregar una **palabra intrusa** que tenga baja probabilidad en el tópico seleccionado y alta probabilidad en otro tópico
- Las 6 palabras se reordenan aleatoriamente y se las muestra a personas para que identifiquen a la **palabra intrusa**

<b>Topic 1</b>	floppy	alphabet	computer	processor	memory	disk
<b>Topic 2</b>	molecule	education	study	university	school	student
<b>Topic 3</b>	islands	island	bird	coast	portuguese	mainland

# Métricas basadas en Human Judgments

## Topic intrusion

- Dado un documento, tomar el título, las primeras oraciones y los 3 tópicos con mayor probabilidad.
- Agregar un **tópico intruso** que tenga baja probabilidad en el documento seleccionado
- Las 4 palabras se reordenan aleatoriamente y se los muestra a personas para que identifiquen al **tópico intruso**

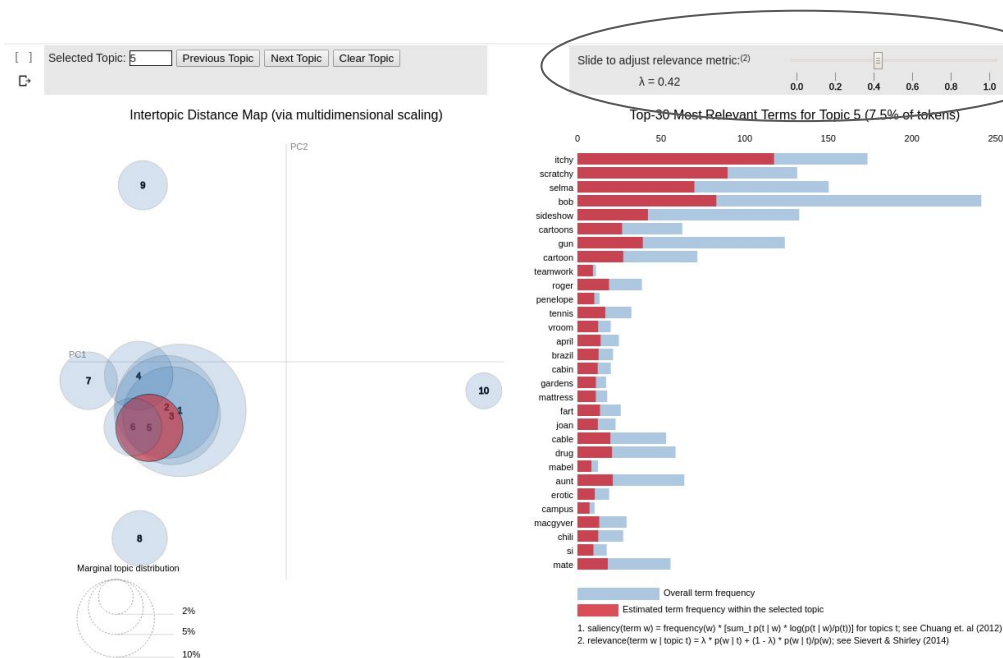
### Douglas Richard Hofstadter

Douglas Richard Hofstadter (born February 15, 1945) is an American professor of cognitive science whose research focuses on the sense of self in relation to the external world, consciousness, analogy-making, artistic creation, literary translation, and discovery in mathematics and physics. Hofstadter's book *Gödel, Escher, Bach: An Eternal Golden Braid*, first published in 1979, won both the Pulitzer Prize for general non-fiction and a National Book Award (at that time called The American Book Award) for Science. His 2007 book *I Am a Strange Loop* won the Los Angeles Times Book Prize for Science and Technology.

education	study	university	school	student	learn
human	life	scientific	science	scientist	lab
play	role	good	actor	star	career
write	work	book	publish	life	friend

# Word *relevance*: LDAvis

■ Frecuencia de la palabra en el corpus  
■ Frecuencia estimada de la palabra en el tópic



Ordenado por word  
*relevance*

Relevance of word  $w$  in topic  $k$ :

$$r(w|k) = \lambda \log(P(w|k)) + (1 - \lambda) \log\left(\frac{P(w|k)}{P(w)}\right)$$

# Gender Stereotypes in Argentine Magazines (... in progress)

Diego Kozlowski<sup>1</sup>, Gabriela Lozano<sup>2</sup>, Fernando Gonzalez<sup>1</sup>, Alfredo Rolla<sup>1</sup>, Jorge Federico Cubells<sup>1</sup>, Edgar Altszyler<sup>1,3</sup>

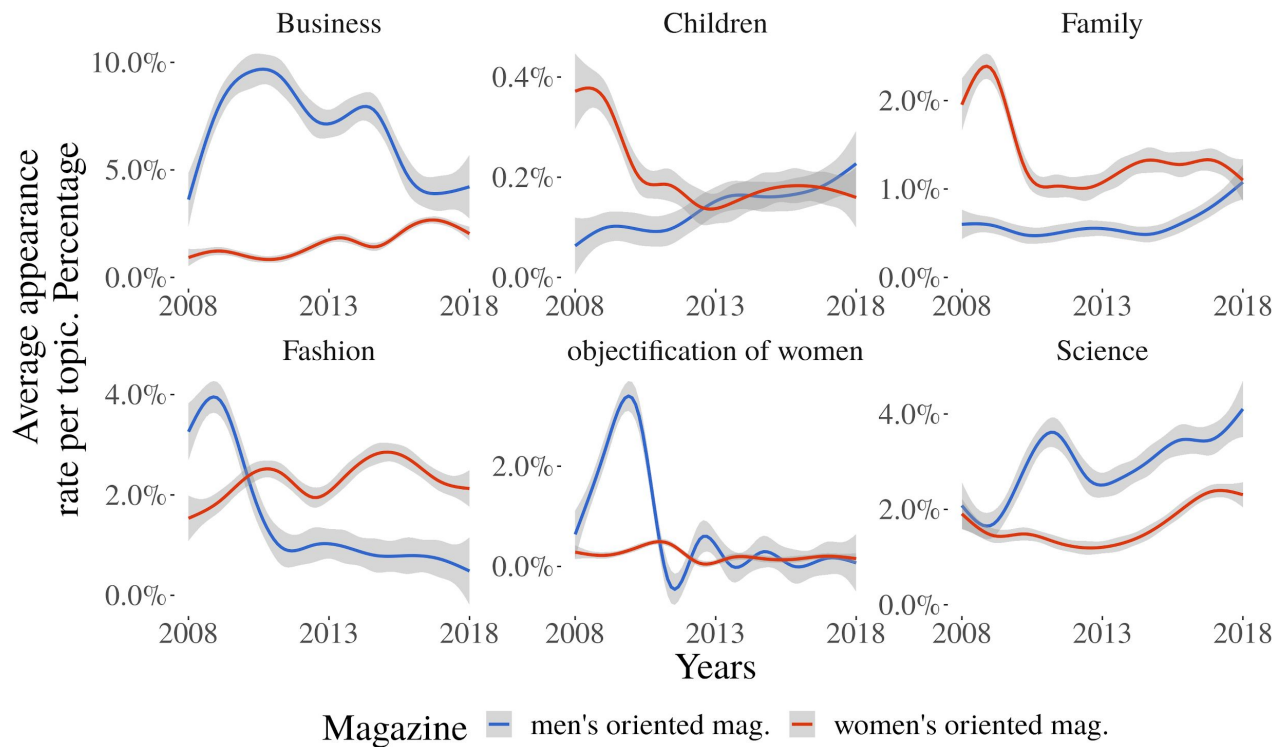


Table 1: Selected Topics

Topic id	assigned Tag	Top 10 words
1	Women	natalia <u>hot</u> ana emma romina <u>versus</u> diez ( <i>ten</i> ) camilo <u>morochas</u>
	objectification	( <i>brunettes</i> ) mega
4	Business	<u>empresa</u> ( <i>company</i> ) redes ( <i>networks</i> ) sistema ( <i>system</i> ) comprar ( <i>to buy</i> ) productos ( <i>products</i> ) mercado ( <i>markets</i> ) traves ( <i>crossing</i> ) tecnologia ( <i>technology</i> ) permite ( <i>it allows</i> ) desarrollo ( <i>development</i> )
7	Children	niños ( <i>kids</i> ) adultos ( <i>adult</i> ) educativo ( <i>educative</i> ) colegio ( <i>school</i> ) chiquito ( <i>tiny</i> ) padre ( <i>father</i> ) secuestro ( <i>kidnap</i> ) change sauna pegote ( <i>goop</i> )
21	Fashion	moda ( <i>fashion</i> ) diseño ( <i>design</i> ) estilo ( <i>style</i> ) marca ( <i>brand</i> ) colección ( <i>collection</i> ) ropa ( <i>cloth</i> ) tendencia ( <i>trend</i> ) prendas ( <i>garments</i> ) rosa ( <i>pink</i> ) zapatillas ( <i>sneakers</i> )
50	Family	hijos ( <i>children</i> ) madre ( <i>mother</i> ) mama ( <i>mom</i> ) padre ( <i>father</i> ) bebe ( <i>baby</i> ) familia ( <i>family</i> ) papa ( <i>dad</i> ) embarazo ( <i>pregnancy</i> ) regalo ( <i>gifts</i> ) años ( <i>years</i> )
82	Science	estudio ( <i>study</i> ) problema ( <i>problem</i> ) trabajo ( <i>work</i> ) explica ( <i>explains</i> ) ley ( <i>law</i> ) medico ( <i>medic</i> ) social ( <i>social</i> ) generar ( <i>generates</i> ) desarrollo ( <i>develops</i> ) investigar ( <i>research</i> )



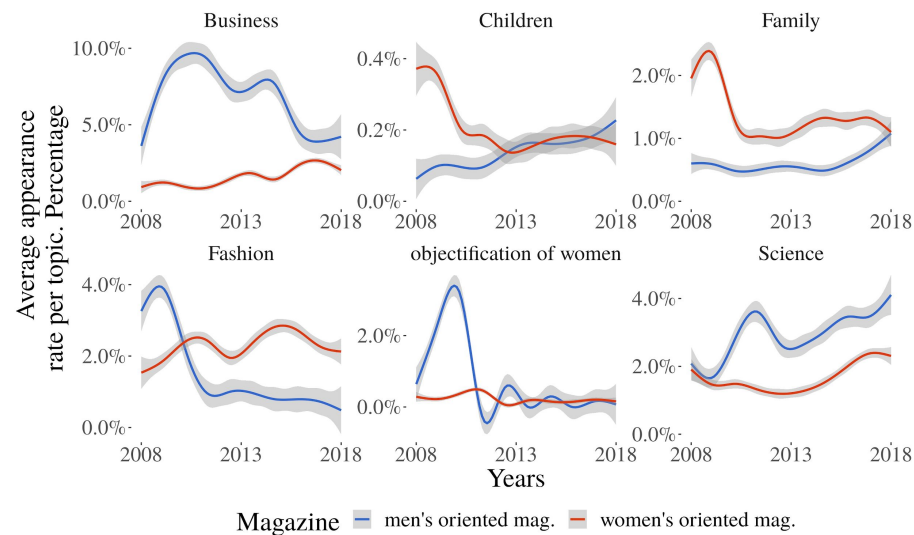
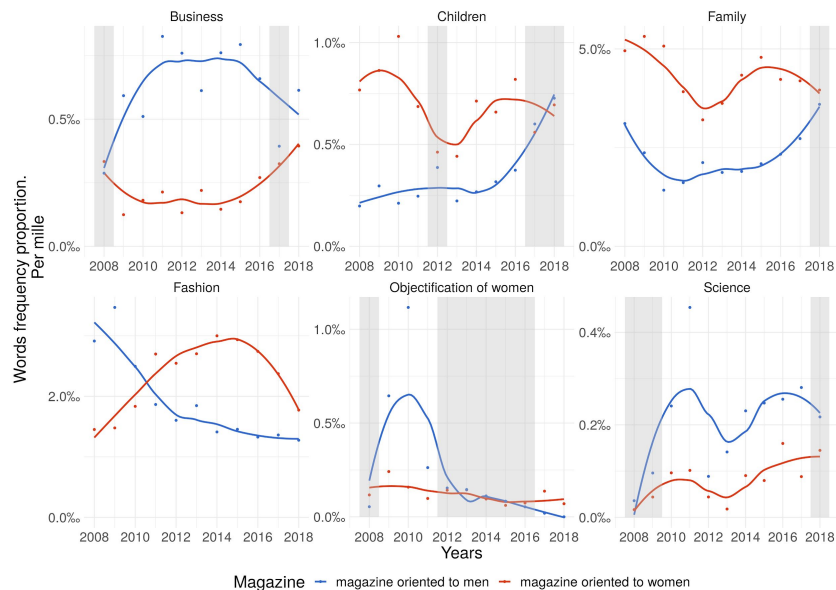
# Evolución de tópicos por revista



Seleccionamos subconjunto de palabras no ambiguas que representen los tópicos

Etiqueta	palabras
Objetivación de la mujer	hot, morocha
Negocio	empresa
Niños	niños, adultos, colegio
Moda	diseño, estilo, ropa
Familia	hijos, madre, mama, padre, papa, bebe, familia
Ciencia	ciencia

# Seleccionamos subconjunto de palabras no ambiguas que representen los tópicos



**FIN**