

## **Módulo 9**

# Procesamiento de Lenguaje Natural o Minería de textos

*Mtro. Luis Enrique Argota Vega*



# Tema 6: Aprendizaje no Supervisado para análisis de textos

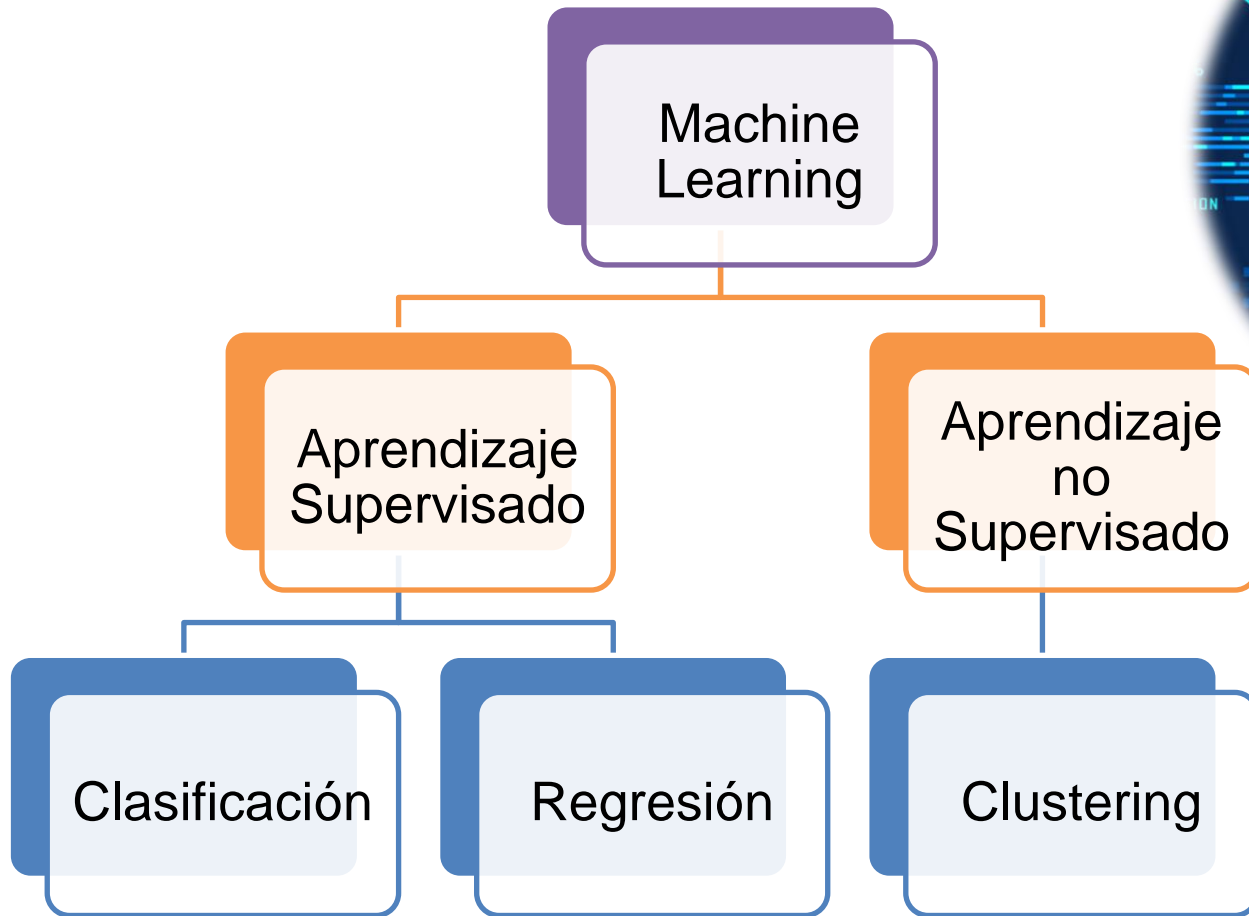
## Objetivo

El participante identificará la técnica de modelado de tópicos para la identificación y análisis de temas en una colección de textos, apoyado en el enfoque de modelado Latent Dirichlet Allocation (LDA) y su implementación en Python.

# Contenido

1. Modelado de tópicos: LDA
2. Agrupamiento de documentos

# Introducción



# Introducci3



Fuente: <http://www.aic.uva.es/cuentapalabras/topic-modeling.html#topic-modeling->



## Examine las nubes de palabras



## *Ritos de muerte*

Almudena Grandes  
**LAS EDADES DE LULÚ**

colección andanzas

Antonio  
**Muñoz Molina**

Ardor guerrero

**TRAFALGAR**

Benito Pérez Galdós

nivola

Las cuatro novelas procesadas y representadas en las cuatro nubes de palabras de la figura son:

1. Ritos de muerte, de Alicia Gimenez-Bartlet (1996),
2. Las edades de Lulú, de Almudena Grandes (1989),
3. Trafalgar, de Benito Pérez Galdós (1873) y
4. Ardor guerrero, de Antonio Muñoz Molina (1995).



# ¿Qué es modelado de tópicos?

- ✓ Consiste en identificar tópicos o temas en textos.
- ✓ Realiza un análisis de lo que hay en una colección de texto.
- ✓ Se asume que un documento es una mezcla de temas y, por otra parte, los temas se representan como una distribución de palabras.
- ✓ Un tópico en el contexto de modelado de tópicos, es una distribución de probabilidades de palabras para un conjunto, e indica la probabilidad que una palabra aparezca en un documento sobre un tópico en particular.



# Enfoques de modelado de temas

- ✓ Análisis semántico latente probabilístico (PLSA) [Hoffman '99]
- ✓ **Latent Dirichlet Allocation (LDA)** [Blei, Ng y Jordan, '03]
- ✓ Basada en aprendizaje profundo (LDA2VEC) [Moody, '16]

El modelado de Latent Dirichlet Allocation (LDA) asume que los documentos se producen a partir de una mezcla de temas. Esos temas luego generan palabras basadas en su distribución de probabilidad. Dado un conjunto de datos de documentos, LDA realiza un seguimiento e intenta averiguar qué temas crearían esos documentos en primer lugar.

## Topics (Temas)

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  
...

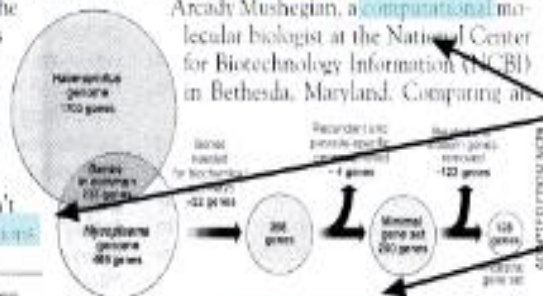
## Documents (Documentos)

### 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 of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a **scientific 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

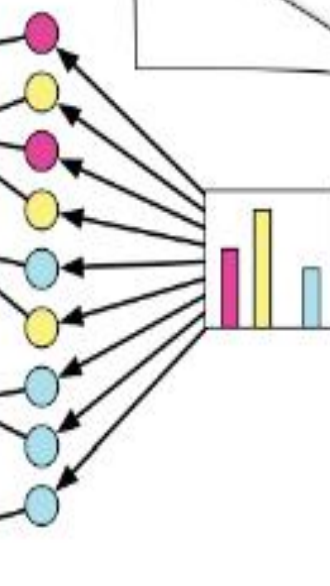


\* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

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

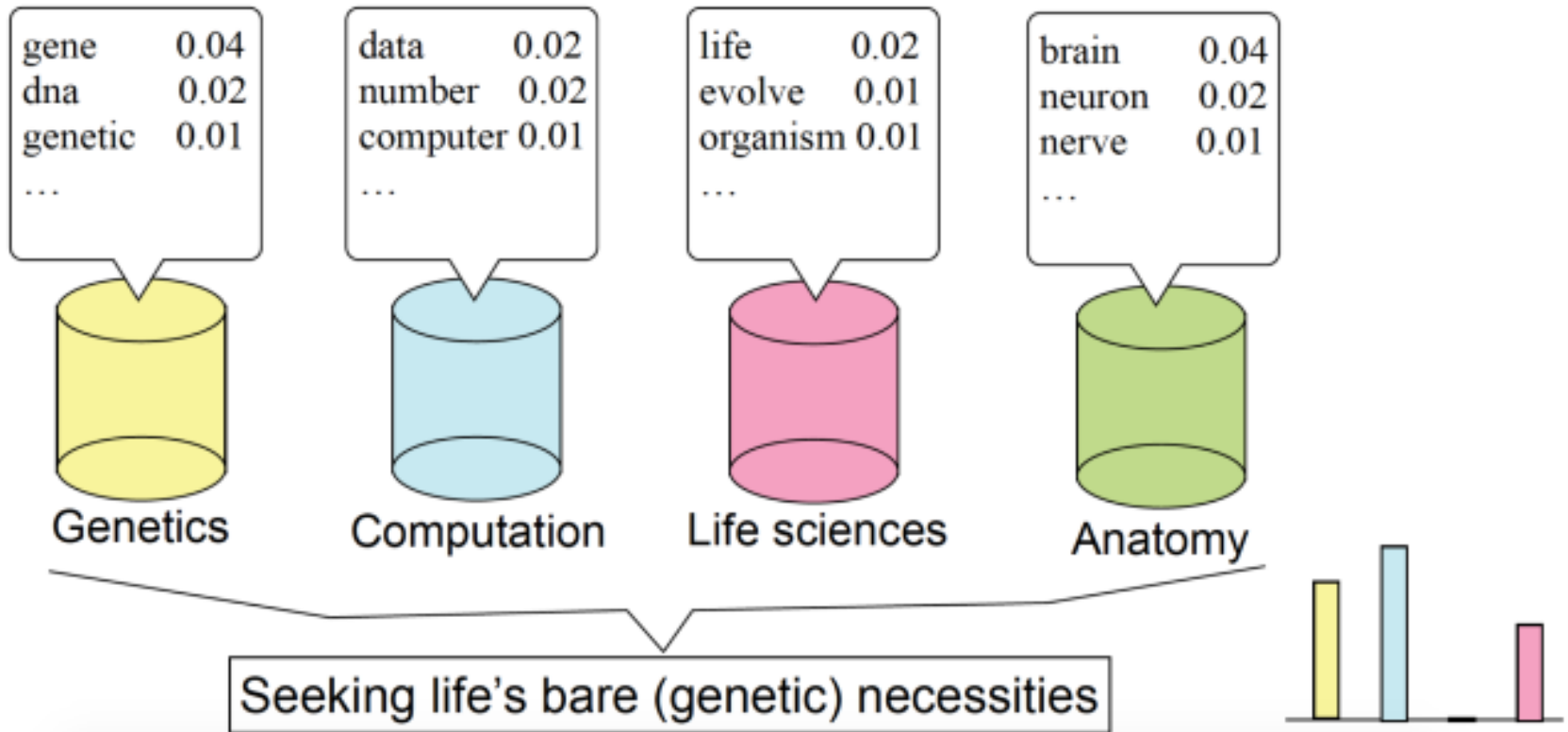
SCIENCE • VOL. 272 • 24 MAY 1996

## Topic proportions and assignments (Asignación y proporción de temas)



Fuente: [http://opac.pucv.cl/pucv\\_txt/txt-5000/UCD5100\\_01.pdf](http://opac.pucv.cl/pucv_txt/txt-5000/UCD5100_01.pdf)

# Intuición: documentos como una mezcla de tópicos



# LDA: Latent Dirichlet Allocation

	$w_1$	$w_2$	$w_3$	$w_m$
$D_1$	0	2	1	3
$D_2$	1	4	0	0
$D_3$	0	2	3	1
$D_n$	1	1	3	0

a. Matriz de Término Documento

	$K_1$	$K_2$	$K_3$	$K_n$
$D_1$	1	0	0	1
$D_2$	1	1	0	0
$D_3$	1	0	0	1
$D_n$	1	0	1	0

b. M1: Matriz de temas de documentos

	$w_1$	$w_2$	$w_3$	$w_m$
$K_1$	0	1	1	1
$K_2$	1	1	1	0
$K_3$	1	0	0	1
$K_n$	1	1	0	0

c. M2: Matriz de temas con dimensiones

Fig. Factorización matricial del LDA

- Convierte la Matriz de Término del Documento (figura a) en dos matrices de dimensiones inferiores:  $M1$  y  $M2$ , utilizando técnicas de muestreo para mejorar estas matrices.
    - $M1$  es una matriz de temas de documentos de dimensión  $(N, K)$  (figura b)
    - $M2$  es una matriz de temas de dimensión  $(K, M)$  (figura c)
- donde  $N$  es el número de documentos,  $K$  es el número de temas y  $M$  es el tamaño del vocabulario.



# LDA: Latent Dirichlet Allocation

- Itera a través de cada palabra " $w$ " para cada documento " $D$ " e intenta ajustar el tema actual. Se asigna un nuevo tema " $K$ " a la palabra " $w$ " con una probabilidad  $P$  que es producto de dos probabilidades  $p_1$  y  $p_2$ . Para cada tema, se calculan dos probabilidades  $p_1$  y  $p_2$ .
  - $p_1 = p\left(\frac{T}{D}\right)$ : la proporción de palabras en el documento y que actualmente están asignadas al tema  $T$ .
  - $p_2 = p\left(\frac{w}{T}\right)$ : la proporción de asignaciones al tema  $t$  sobre todos los documentos que provienen de esta palabra  $w$ .
- Después de varias iteraciones, se logra un estado estable donde el tema del documento y las distribuciones de los términos del tema son bastante buenos.

# Modelado de tópicos en la práctica

## ¿Cuántos temas?

- Encontrar o incluso adivinar la cantidad de temas es difícil

## Interpretar tópicos

- Los temas son solo distribuciones de palabras.
- Dar sentido a las palabras/generar etiquetas es subjetivo

# Modelado de tópicos en la práctica

¿Por qué es útil el modelado de tópicos?

- Clasificación de texto
- Sistemas de recomendación
- Descubrir temas en textos

# Trabajando con LDA en Python

- ✓ Muchos paquetes disponibles, como: gensim, sklearn
- ✓ Preprocesamiento de texto
  - Tokenizar, normalizar (minúsculas)
  - Eliminar palabras cerradas
  - Stemming / Lematizar
- ✓ Convertir documentos tokenizados a una matriz de termino-documento
- ✓ Construir modelos LDA en la matriz termino-documento

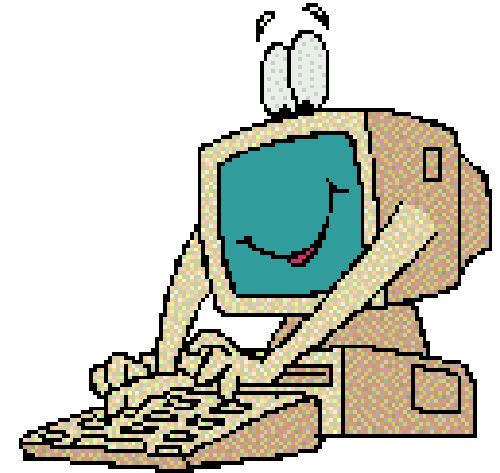
`conda install -c anaconda gensim`

`conda install -c anaconda scikit-learn`

`conda install -c anaconda nltk`



# Práctica



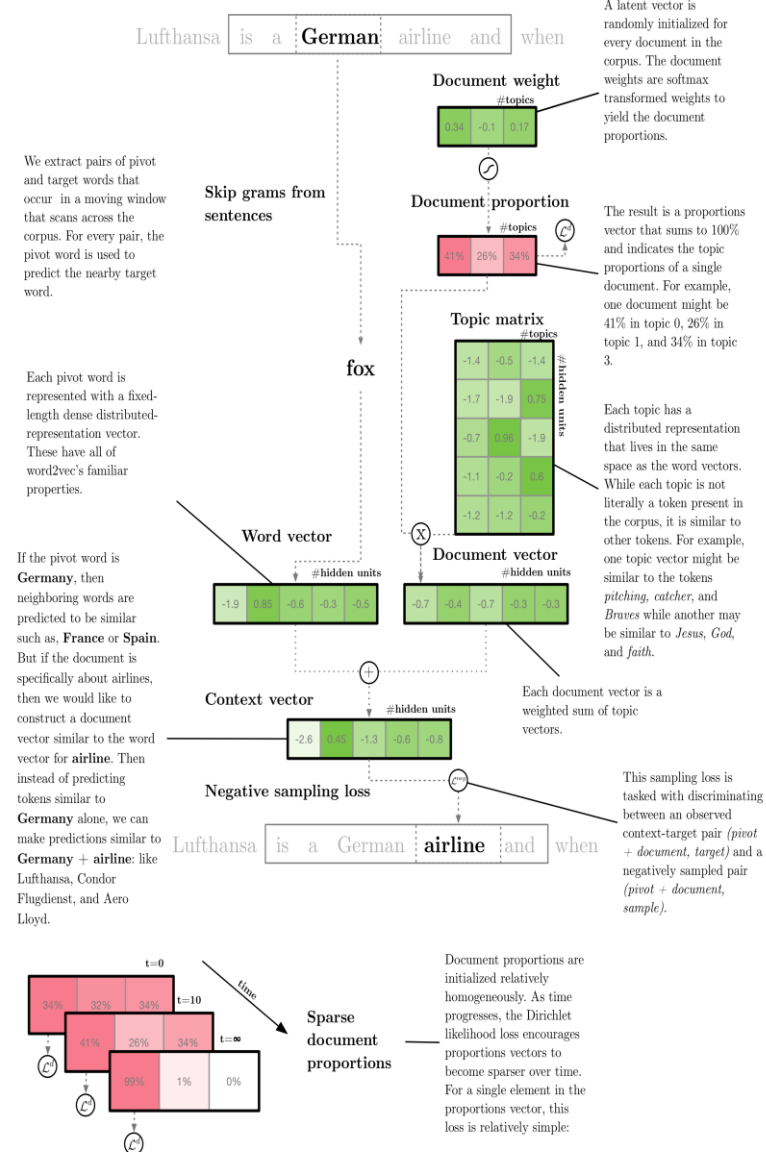
*Ejercicio6(es)-Aprendizaje no Supervisado.ipynb*

# Tendencias

**LDA2VEC:** Aprende simultáneamente representaciones de tópicos y representaciones de documentos también. El modelo lda2vec intenta mezclar las mejores partes de word2vec y LDA en un solo marco

<https://arxiv.org/abs/1605.02019>

Código de experimentos, software de investigación en Python: <https://github.com/cemoody/lda2vec>



# Conclusiones

- El modelado de tópicos es una herramienta exploratoria, frecuentemente utilizada para extracción de textos.
- LDA es un modelo generativo, utilizado extensivamente para modelar grandes corpus de texto.
- LDA también se puede utilizar como una técnica de selección de características, para clasificación de textos y otras tareas.



# Referencias

- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. the Journal of machine Learning research, 3, 993-1022.
- Ponweiser, M. (2012). Latent Dirichlet allocation in R. Obtenido de <https://epub.wu.ac.at/3558/>
- Priyantina, R., & Sarno, R. (2019). Sentiment Analysis of Hotel Reviews Using Latent Dirichlet Allocation, Semantic Similarity and LSTM. International Journal of Intelligent Engineering and Systems, 12(4), 142-155. Obtenido de <http://www.inass.org/2019/2019083114.pdf>



# Contacto

Luis Enrique Argota Vega

*Máster en Ciencia e Ingeniería de la Computación*

luiso91@gmx.com

Tels: 5578050838

Redes sociales:



<https://cutt.ly/ifPyTEH>



<https://cutt.ly/WfPtYZz>