

# Topic Modeling

...

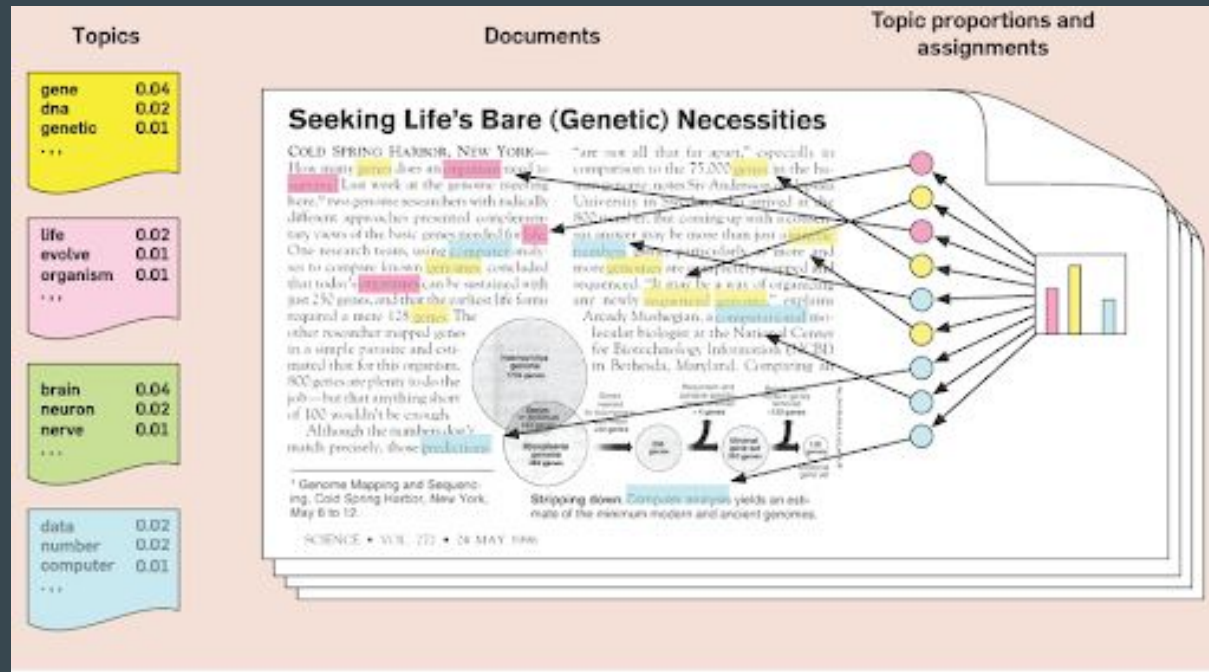
KeepCoding - Bootcamp de Big Data & Machine Learning

# Introducción

- Conjunto de técnicas para **descubrir estructuras latentes semánticas comunes** (topics / temas) en un conjunto de documentos
- Surge del problema: ¿cómo puedo **sintetizar la información** en una colección grande de documentos con información semi estructurada?
- Latent Dirichlet Allocation (**LDA**)

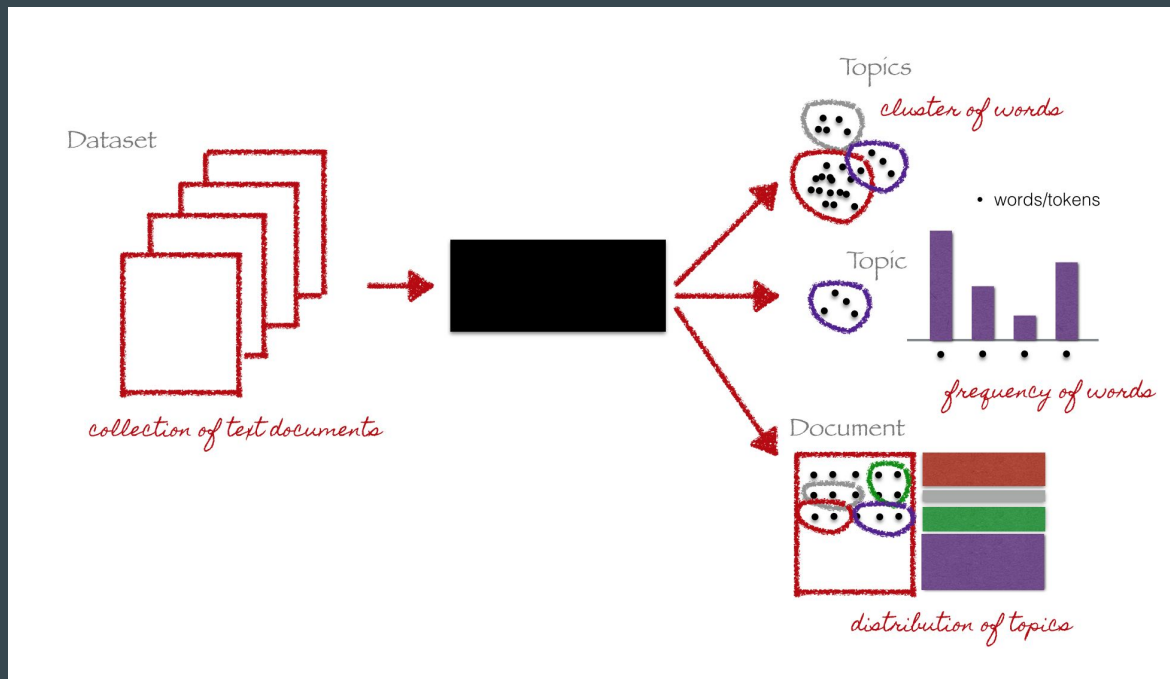
# Introducción

## Idea principal



# Introducción

## Idea principal



# LDA (Latent Dirichlet Allocation)

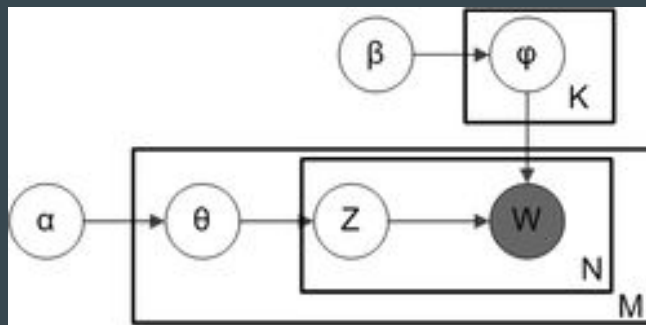
- Modelo probabilístico probabilístico
- Aprendizaje **no supervisado** (no tenemos información a priori de los posibles topics que hay o, al menos, no están etiquetados)
- Asume que
  - **Documentos con topics similares usarán palabras similares**
  - Los **documentos** están **compuestos por un conjunto de topics** (que siguen una determinada distribución)
  - Los **topics** están **compuestos por un conjunto de palabras** (que siguen una determinada distribución)

# LDA (Latent Dirichlet Allocation)

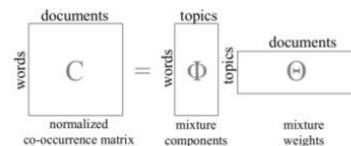
- Debe fijarse el vocabulario al inicio
- Conviene preprocesar. En este caso, eliminar stop words suele arrojar mejores resultados
- Representación de bag-of-words
- Debemos definir el número de topics que queremos que extraiga (similar al k-means)

# LDA - Parámetros

- $M$  - número de documentos
- $N$  - número de palabras en el documento  $M$
- $\alpha$  - Factor de densidad de doc-topic
- $\beta$  - Factor de densidad de topic-word
- $\theta_i$  - Distribución de topics en el documento  $M$
- $\phi_k$  - Distribución de palabras en el topic  $K$
- $Z_{ij}$  - Topic para la palabra  $j$  en el documento  $i$
- $w_{ij}$  - Palabra (variable observable)

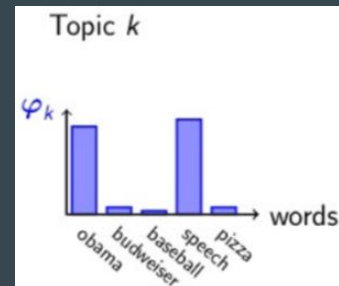
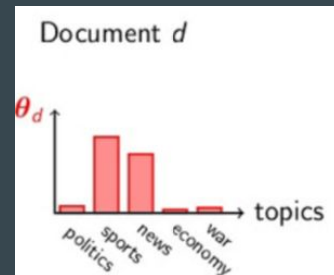


## Matrix Factorization Interpretation of LDA



# LDA - Distribuciones

- Los documentos estarán compuestos por un conjunto de topics (que siguen una determinada distribución)
- Los topics estarán compuestos por un conjunto de palabras (que siguen una determinada distribución)





¡Vamos al lío!