Minería de texto en Twitter

Laura Alonso Alemany
Grupo de Procesamiento del Lenguaje Natural
FAMAF-UNC

Workshop sobre Big Data en Ciencias Económicas FCE-UNC 2019

Redes sociales como indicadores

- Termómetro de tendencias con posible impacto económico
  - Valores de marcas
  - Percepción de oportunidad y de riesgo
  - Movimientos poblacionales, estacionales, etc.
- Identificación de comunidades, actores importantes
- Identificación de temas relevantes

Identificar temas

El lenguaje natural es un **sistema complejo**,
con **propiedades emergentes**

Las palabras son **fenómenos observables** de **causas latentes**

Los temas son las causas que originan las palabras

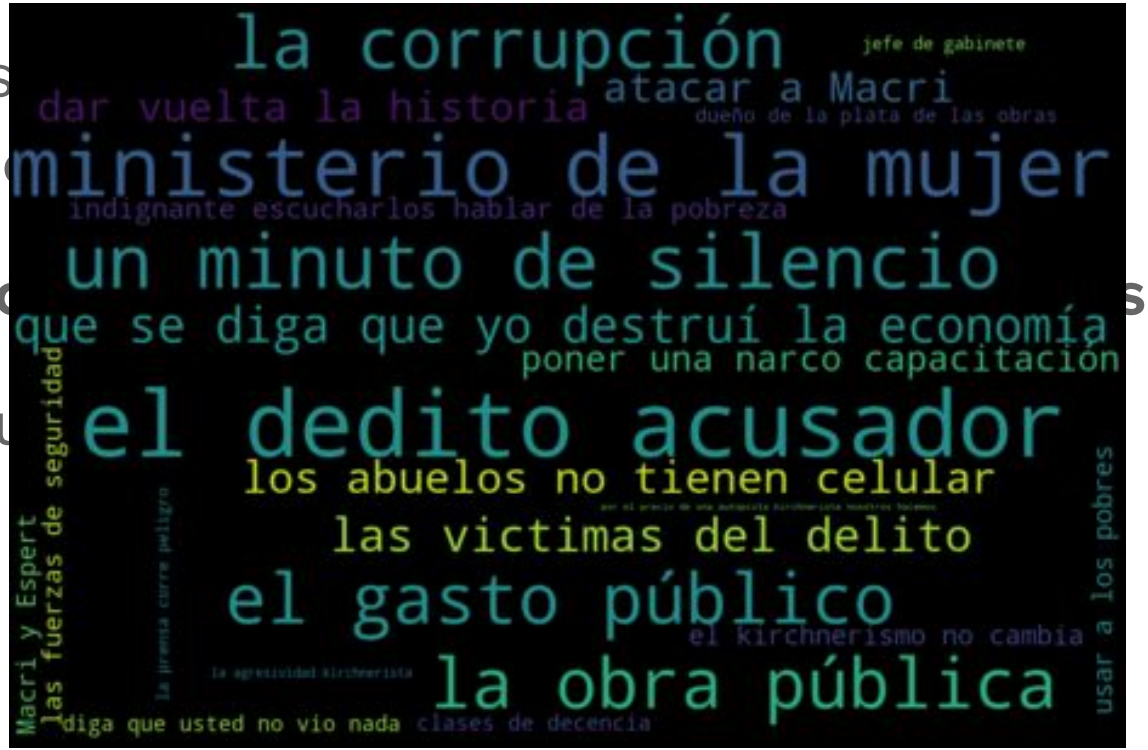
Identificar temas

El lenguaje natural es

con **propiedades em**

Las palabras son **fen** **s**

Los temas son las cau



https://prensaobrera.com/politicas/66076-las-redes-sociales-durante-los-debates-presidenciales

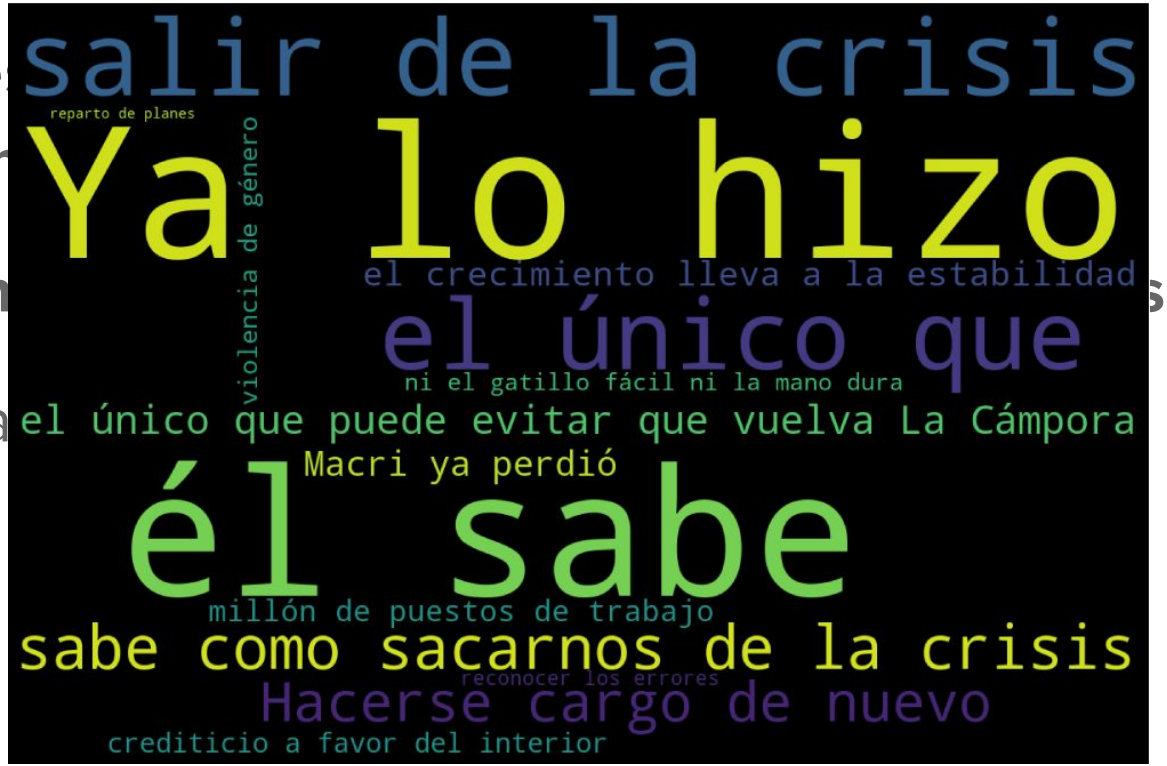
Identificar temas

El lenguaje natural e

con **propiedades em**

Las palabras son **fen** s

Los temas son las ca



https://prensaobrera.com/politicas/66076-las-redes-sociales-durante-los-debates-presidenciales

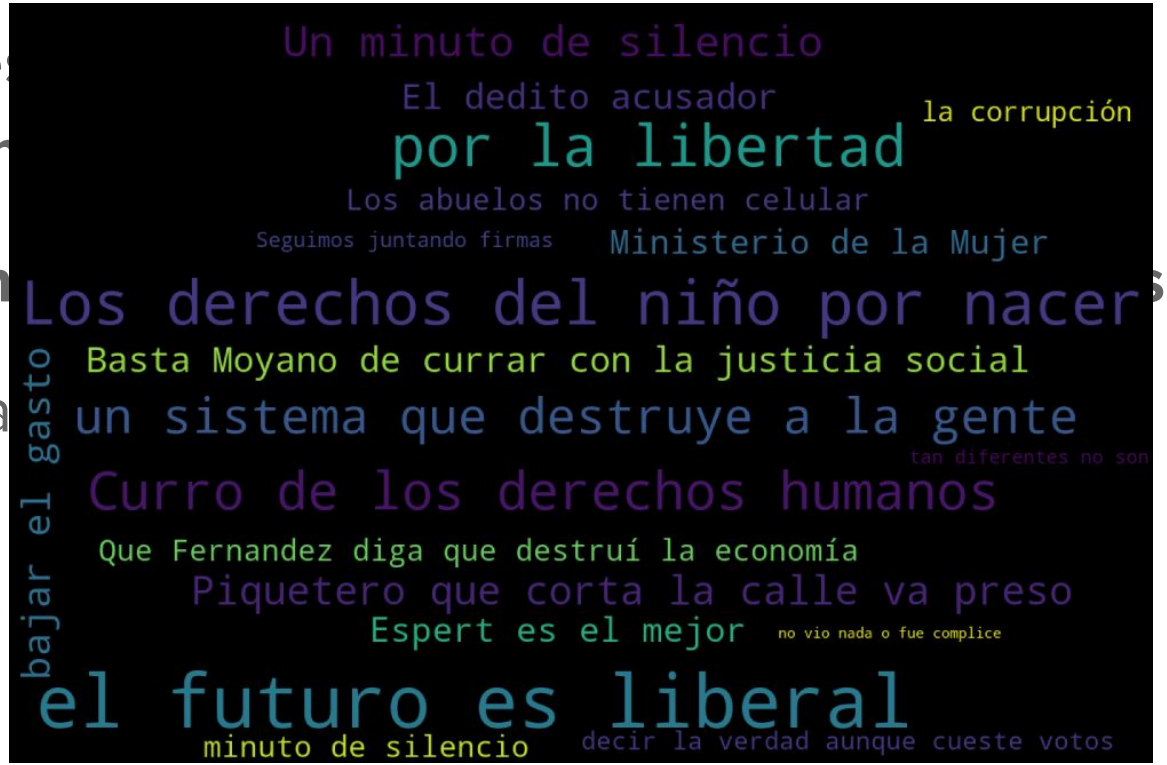
Identificar temas

El lenguaje natural es

con **propiedades em**

Las palabras son **fen**

Los temas son las ca



https://prensaobrera.com/politicas/66076-las-redes-sociales-durante-los-debates-presidenciales

Procesamiento de tweets

Por qué twitter?

- Porque el resto de redes sociales no están disponibles públicamente
- Es fácil obtener una cierta cantidad de datos (API)
- Los metadatos son ricos y podemos estudiar diferentes perspectivas

Texto como insumo de Machine Learning

Convertir textos en vectores

Texto como insumo de Machine Learning

Convertir textos en vectores

1. Cada palabra es una dimensión (*BoW, bag-of-words*)

Texto como insumo de Machine Learning

Convertir textos en vectores

1. Cada palabra es una dimensión (*BoW, bag-of-words*)

   - Doc1: Text mining is to identify useful information.
   - Doc2: Useful information is mined from text.
   - Doc3: Apple is delicious.

	text	information	identify	mining	mined	is	useful	to	from	apple	delicious
Doc1	1	1	1	1	0	1	1	1	0	0	0
Doc2	1	1	0	0	1	1	1	0	1	0	0
Doc3	0	0	0	0	0	1	0	0	0	1	1

Texto como insumo de Machine Learning

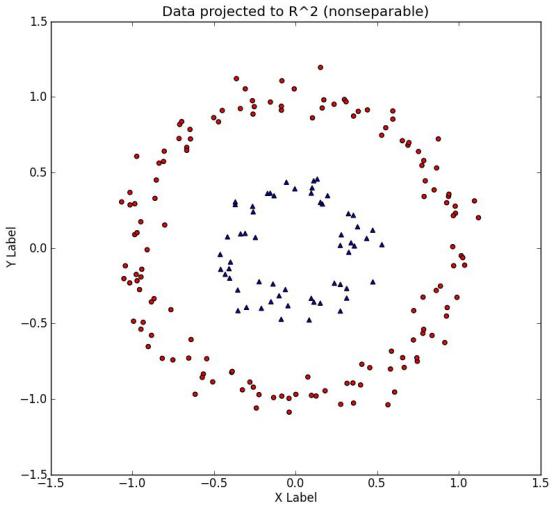
Convertir textos en vectores

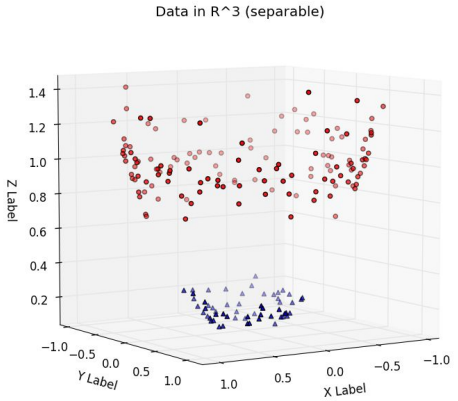
1. Cada palabra es una dimensión (*BoW, bag-of-words*)
2. Se proyecta a otro espacio (*word embeddings*)

Texto como insumo de Machine Learning

Convertir textos en vectores

1. Cada palabra es una dimensión (*BoW, bag-of-words*)
2. Se proyecta a otro espacio (*word embeddings*)





Texto como insumo de Machine Learning

Convertir textos en vectores

1. Cada palabra es una dimensión (*BoW, bag-of-words*)
2. Se proyecta a otro espacio (*word embeddings*)
3. Se encuentran grupos, se clasifica (*clustering, classification*)

Texto como insumo de Machine Learning

Convertir textos en vectores

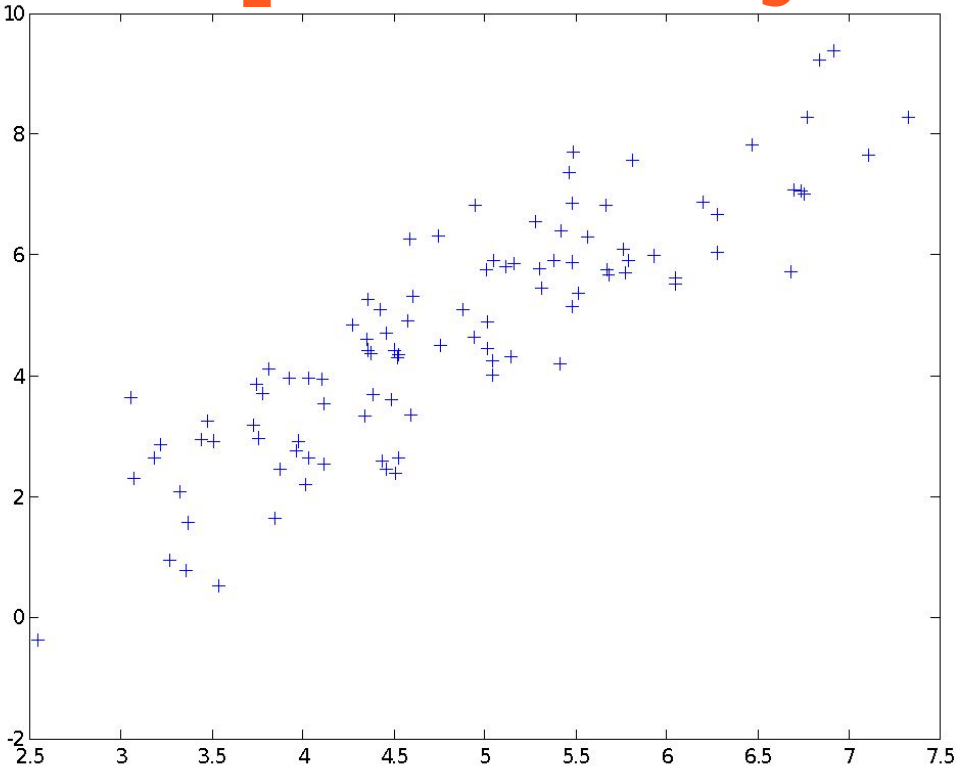
1. Cada palabra es una dimensión (*BoW, bag of words*)
2. Se proyecta a otr[illegible]cio (*w*
3. Se encuentran grupos, [illegible]

pre-proceso:

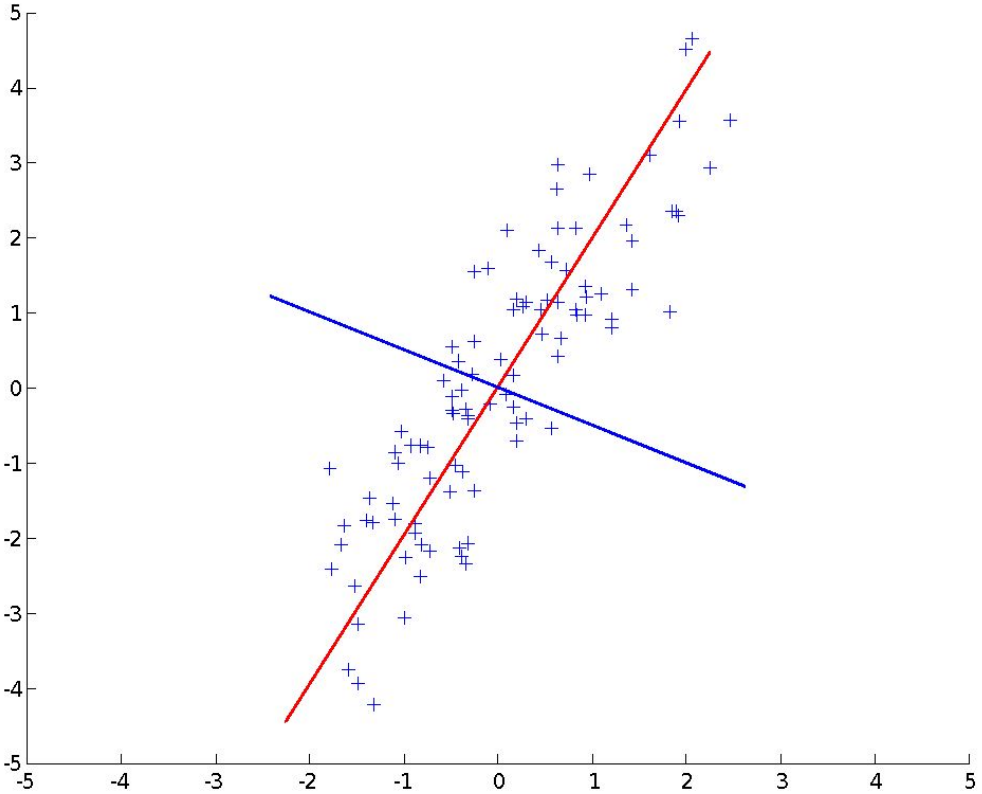
- normalización (*tokenizing*)
- canonicalización lingüística (*lemmatizing*)
- eliminar signos y símbolos
- eliminar muy frecuentes (*stopwords*)
- eliminar poco frecuentes

Proyección a otro espacio

Principal Component Analysis



Principal Component Analysis



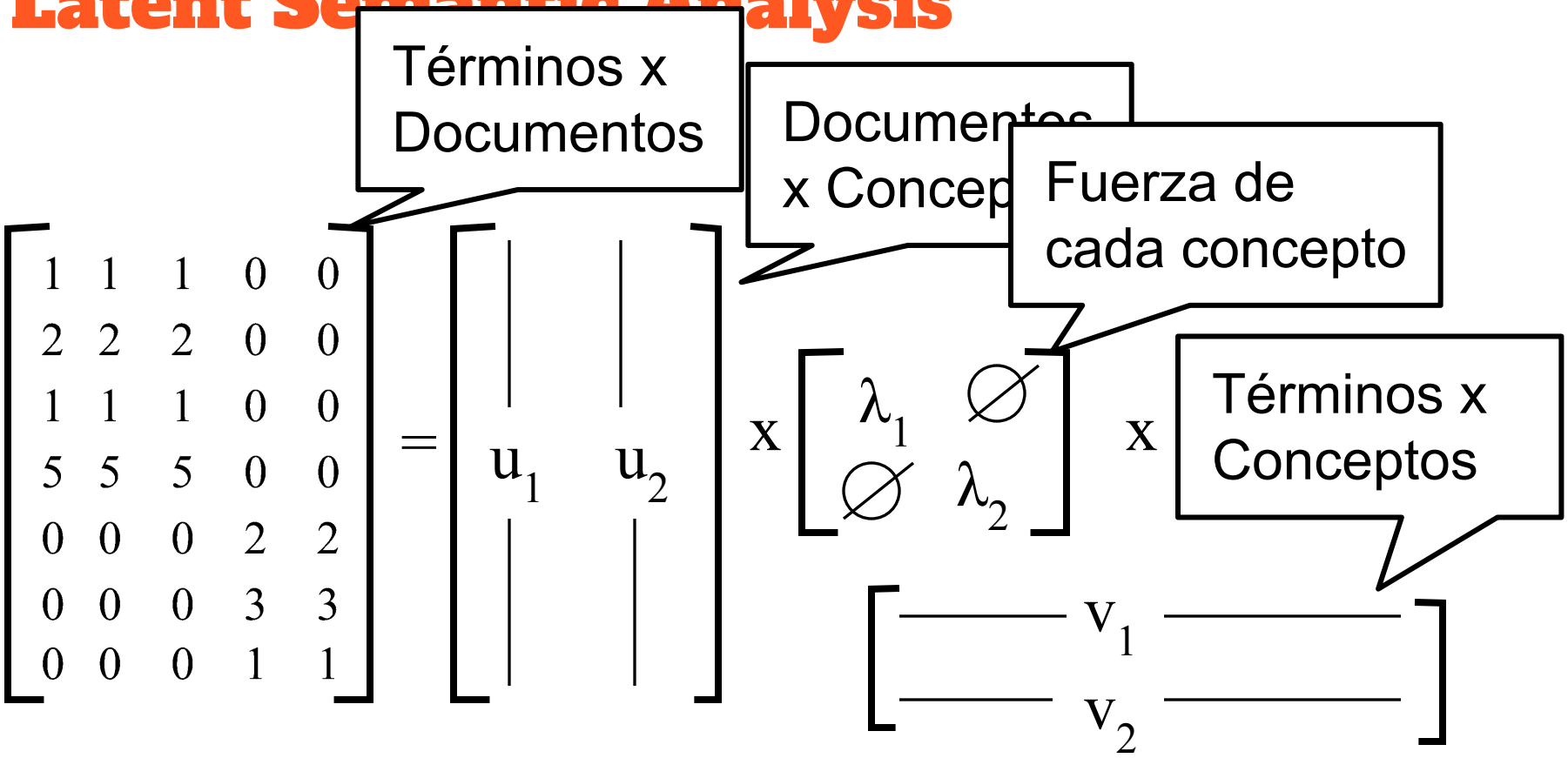
Descomposición en Valores Singulares

Descomponer una matriz en valores singulares (eigenvalues)

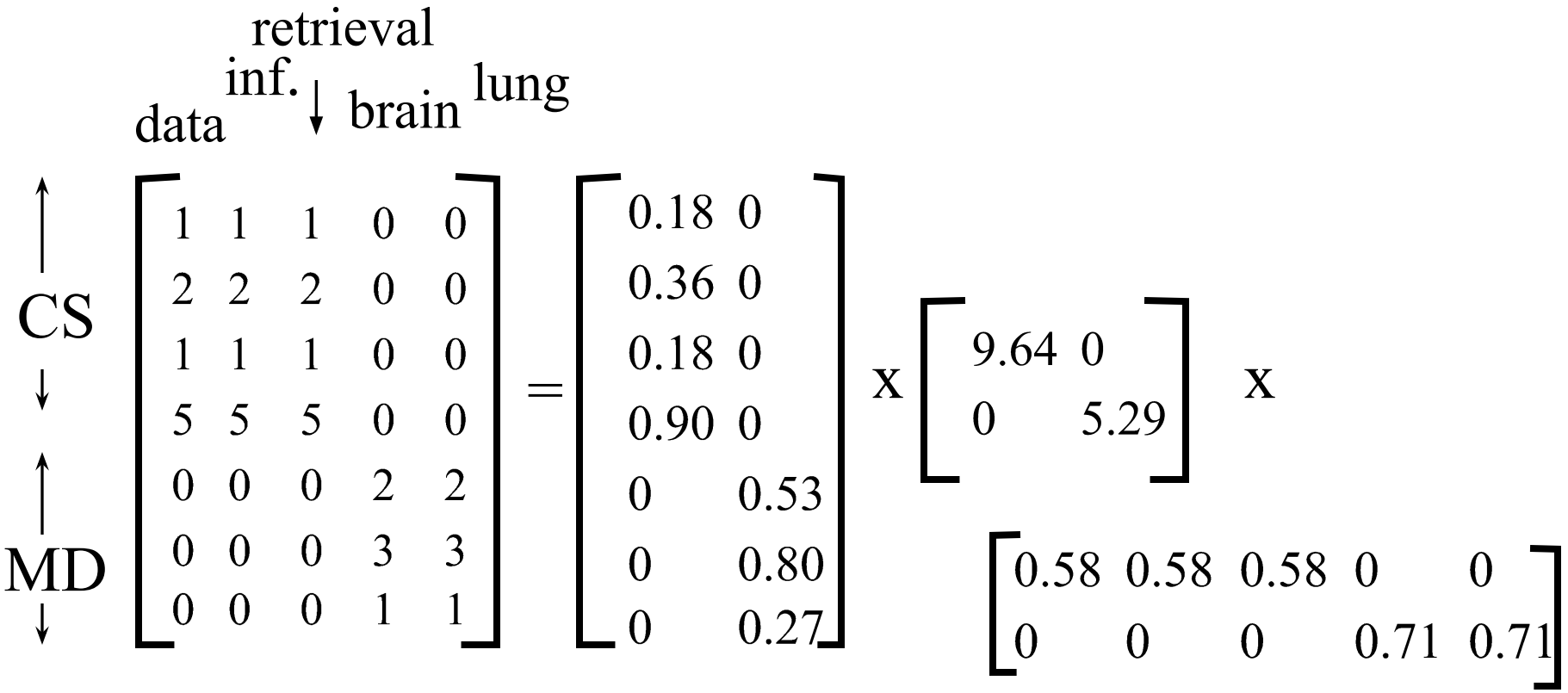
➜ singular value decomposition (SVD)

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix} = \begin{bmatrix} | & | \\ u_1 & u_2 \\ | & | \end{bmatrix} \times \begin{bmatrix} \lambda_1 & \varnothing \\ \varnothing & \lambda_2 \end{bmatrix} \times \begin{bmatrix} — & v_1 & — \\ — & v_2 & — \end{bmatrix}$$

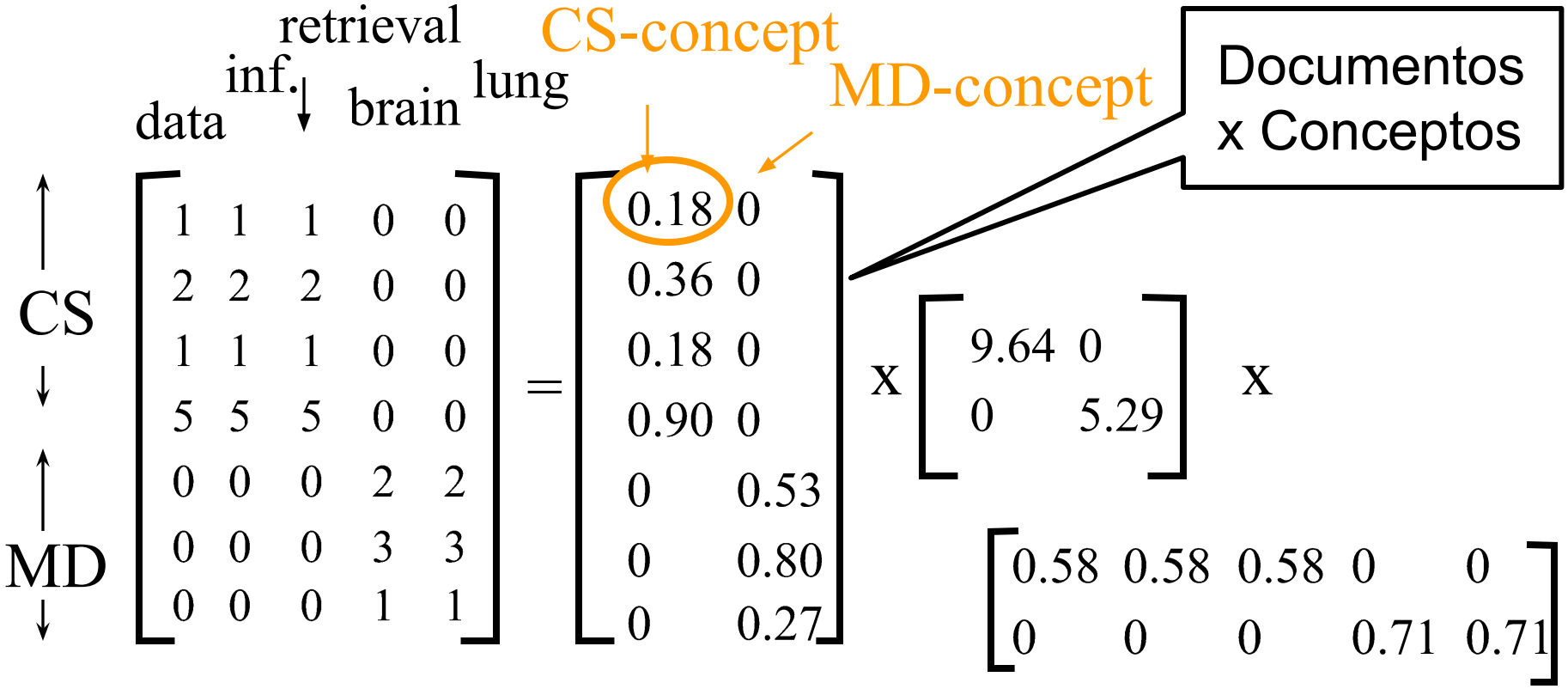
Latent Semantic Analysis



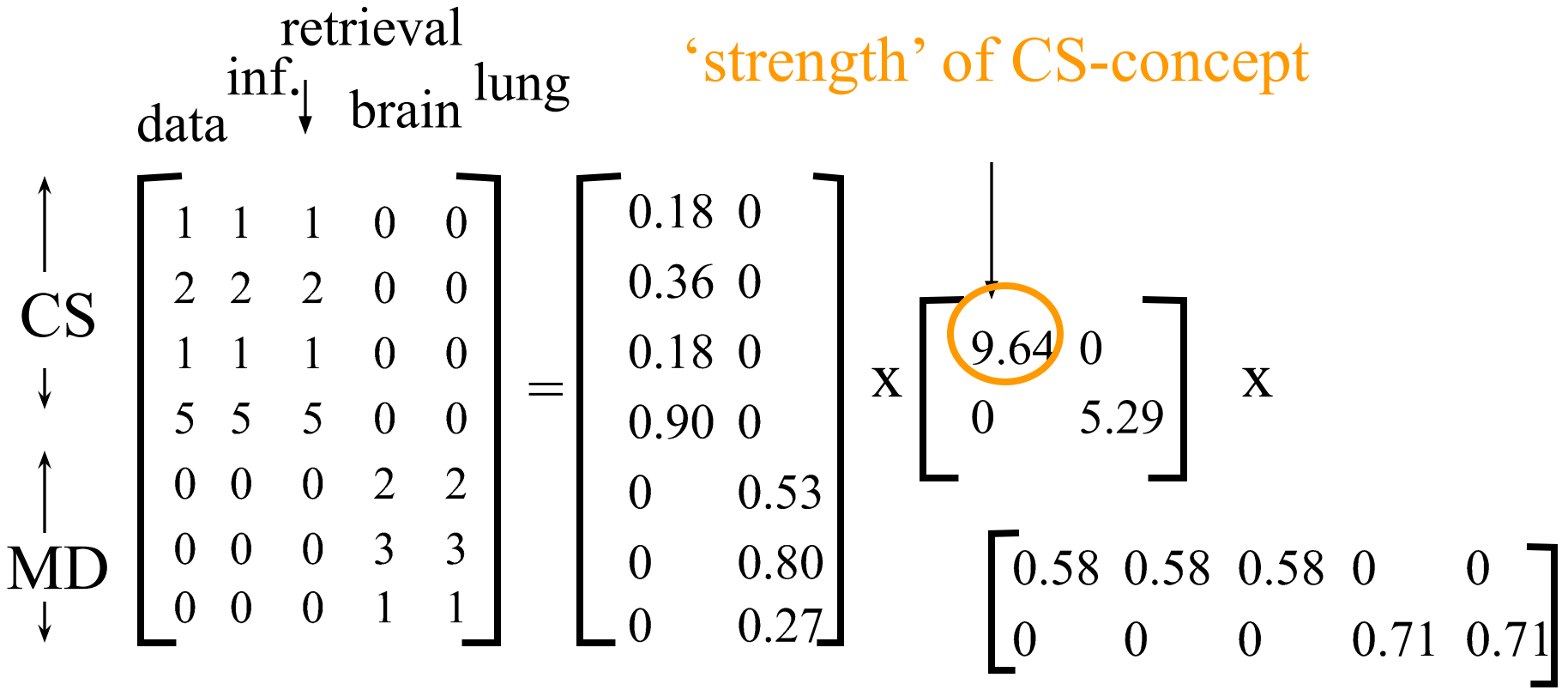
Latent Semantic Analysis



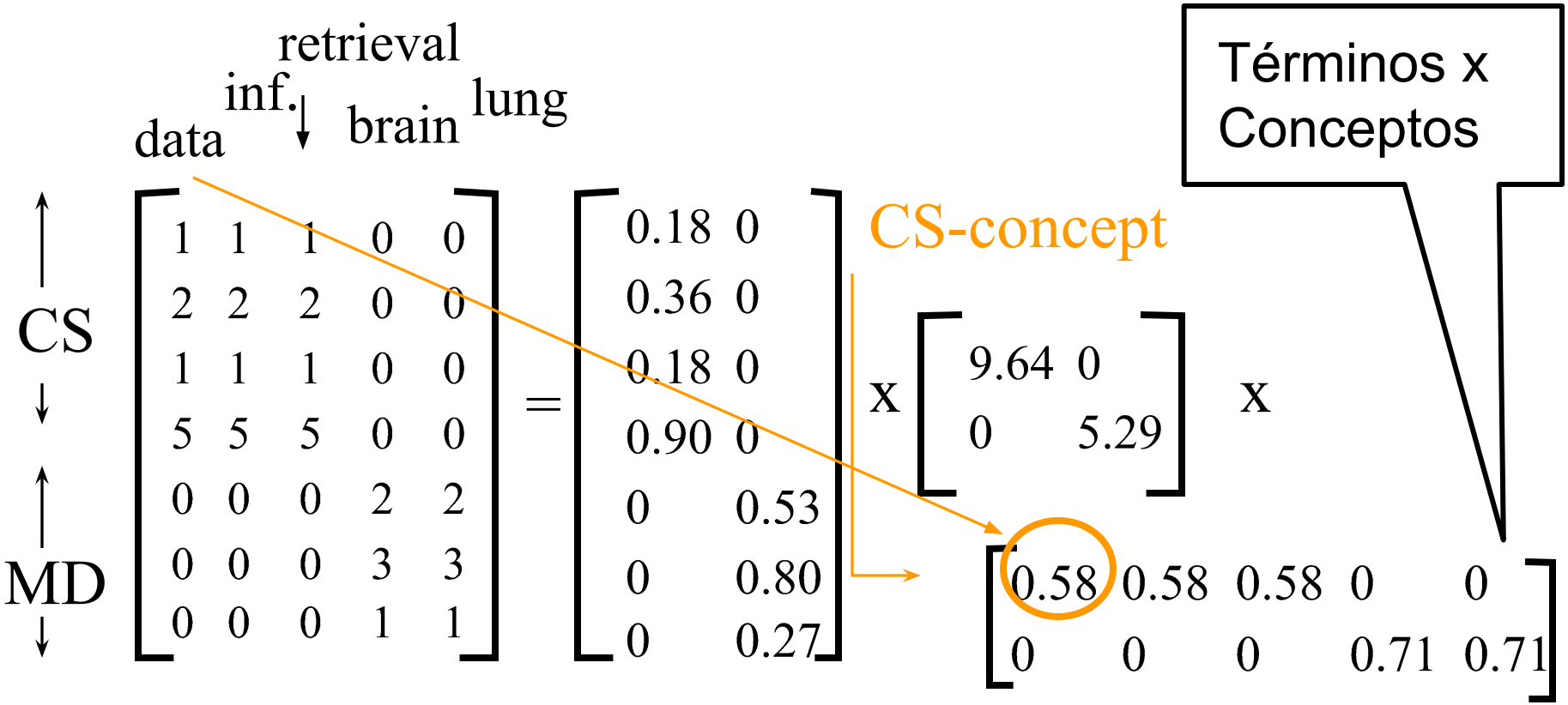
Latent Semantic Analysis



Latent Semantic Analysis



Latent Semantic Analysis



Latent Semantic Analysis: Reducción de dimensionalidad

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 0.18 & 0 \\ 0.36 & 0 \\ 0.18 & 0 \\ 0.90 & 0 \\ 0 & 0.53 \\ 0 & 0.80 \\ 0 & 0.27 \end{bmatrix} \times \begin{bmatrix} 9.64 & 0 \\ 0 & 5.29 \end{bmatrix} \times \begin{bmatrix} 0.58 & 0.58 & 0.58 & 0 & 0 \\ 0 & 0 & 0 & 0.71 & 0.71 \end{bmatrix}$$

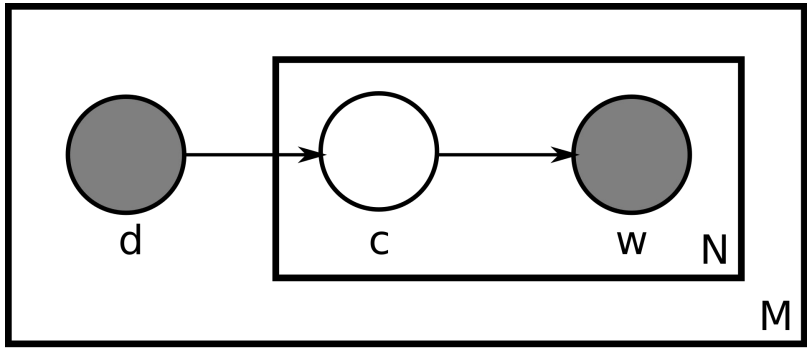
Latent Semantic Analysis: Reducción de dimensionalidad

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix} \sim \begin{bmatrix} 0.18 \\ 0.36 \\ 0.18 \\ 0.90 \\ 0 \\ 0 \\ 0 \end{bmatrix} \times \begin{bmatrix} 9.64 \end{bmatrix} \times \begin{bmatrix} 0.58 & 0.58 & 0.58 & 0 & 0 \end{bmatrix}$$

Detección de tópicos

Probabilistic Latent Semantic Analysis

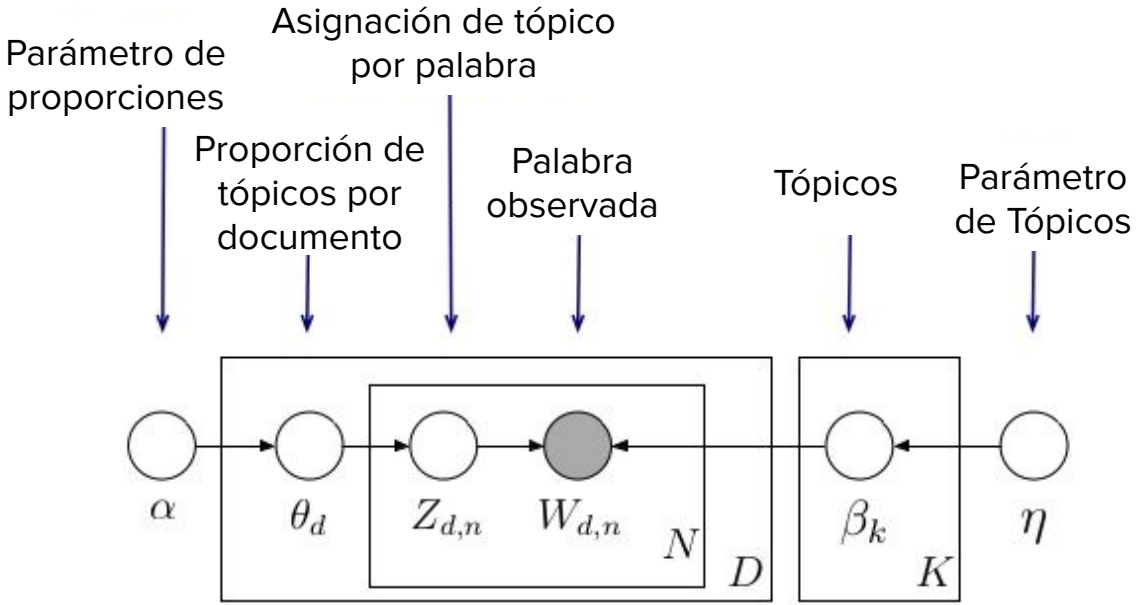
Mezcla de distribuciones multinomiales independientes o clases latentes o tópicos (el n de tópicos es un parámetro)



- **d** es el documento
- **c** es un tópico de la distribución del documento **P(c|d)**
- **w** es una palabra de la distribución de palabras de **c**

Latent Dirichlet Allocation

Mezcla de distribuciones multinomiales con distribución de Dirichlet



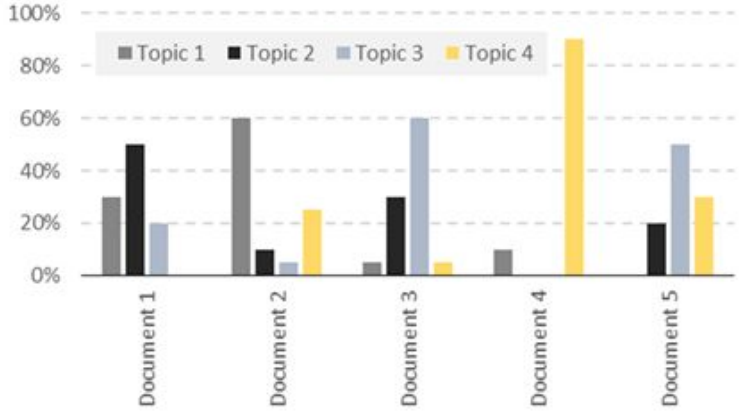
Latent Dirichlet Allocation

Documentos

Inferencia de tópicos

	weight	words
Topic 1	3%	flower
	2%	rose
	1%	plant
...		
Topic 2	2%	company
	1%	wage
	1%	employee

Asignación de tópicos a documentos



¡Manos a la obra!

Usar una notebook de Jupyter

.ipynb

- se ejecuta desde el navegador
- podemos usar mucho código público con comentarios
- se integran visualizaciones

1. bajar Anaconda
2. ejecutar!

https://jupyter-notebook-beginner-guide.readthedocs.io/en/latest/execute.html

Usar una notebook de Jupyter

.ipynb

- se ejecuta desde el navegador
- podemos usar mucho código público con comentarios
- se integran visualizaciones

también pueden usar google colab

1. bajar Anaconda
2. ejecutar!

https://jupyter-notebook-beginner-guide.readthedocs.io/en/latest/execute.html

Pipeline para bajar y preprocesar tweets

https://github.com/pablocelayes/twitter-aborto

https://cs.famaf.unc.edu.ar/~laura/ecobigdata2019.html