

Representación multivariante para modelos LDA de datos textuales

II Congreso de Datos Abiertos y Metadatos

CODAM 2022
DATOS ABIERTOS
EN LA CIENCIA Ecuador

26 al 28 de octubre de 2022 / Modalidad Híbrida



Ing. Ind. Luis Pilacuan B. PhD(c)

Universidad de Guayaquil,
Guayaquil – Ecuador,
luis.pilacuanb@ug.edu.ec

ORGANIZADORES

data-lat.org



AUSPICIANTES

CS&S Code for
Science &
Society

**Chan
Zuckerberg
Initiative**

GORDON AND BETTY
MOORE
FOUNDATION

SUMARIO/CONTENIDO:

INTRODUCCIÓN

- 1.1. Conceptos básicos: Text Mining, Topics Models, Latent Dirichlet Allocation (LDA) , Hj-Biplot**
- 1.2. LDABiplots paquete para Minería de Texto**
- 1.3. Ejemplo Práctico en RStudio**

Octubre, 2022

ORGANIZADORES

data-lat.org



AUSPICIANTES



Introducción

En los últimos años, con el avance tecnológico de la ciencia, la empresa, la academia y la sociedad en general, se ha producido por consecuencia una mayor generación de información. Esta creciente producción de datos la encontramos de manera estructurada como ya veníamos acostumbrados a trabajar, con bases de datos relacionadas, pero ahora también se ha generado una gran cantidad de datos no estructurados, tales como correos electrónicos, redes sociales, imágenes, texto y otros, que precisan de ser extraídos, explorados, analizados, modelados y evaluados; con el fin de tomar decisiones acertadas y para ello se necesitan técnicas especializadas en dichas tareas, que puedan identificar patrones, relaciones o asociaciones existentes en estos datos.



La Estadística Textual trabaja sobre datos que son palabras, provenientes de textos libres (títulos, resúmenes), de encuestas (preguntas abiertas) o de textos integrales.

Los Análisis Textuales implican contar las

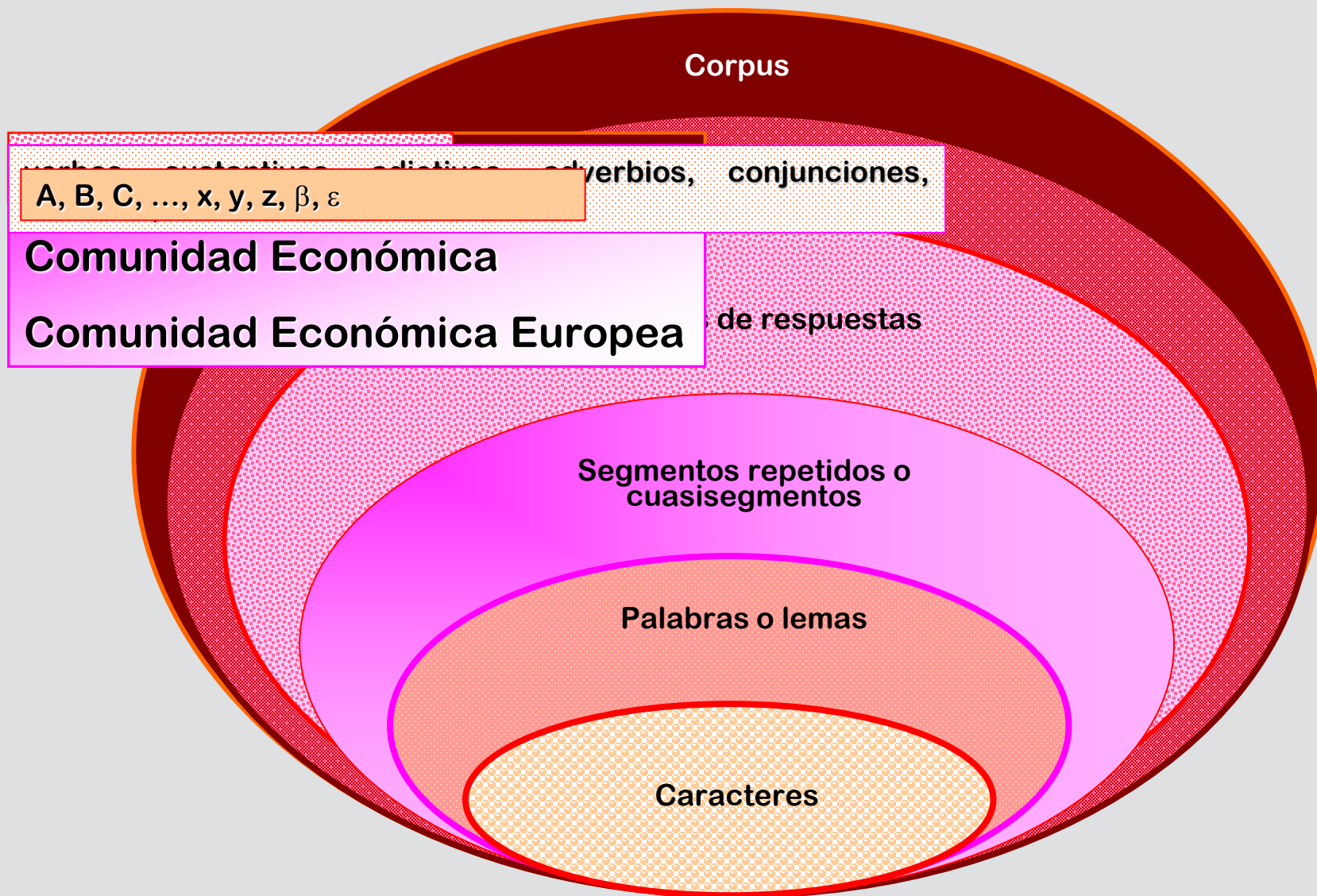
OCURRENCIAS de las **UNIDADES LÉXICAS** básicas en un

CORPUS y operar algún tipo de análisis estadístico a

Colección de uno o más textos, subdivididos en partes codificadas por el usuario.

Usualmente palabras o formas gráficas, lemas, segmentos repetidos, cuasisegmentos, etc.

Número de veces que cada unidad léxica básica se repite en el interior del corpus o en algunas partes de él.



Tomado de Zulaima 2006

ORGANIZADORES

data-lat.org



AUSPICIANTES

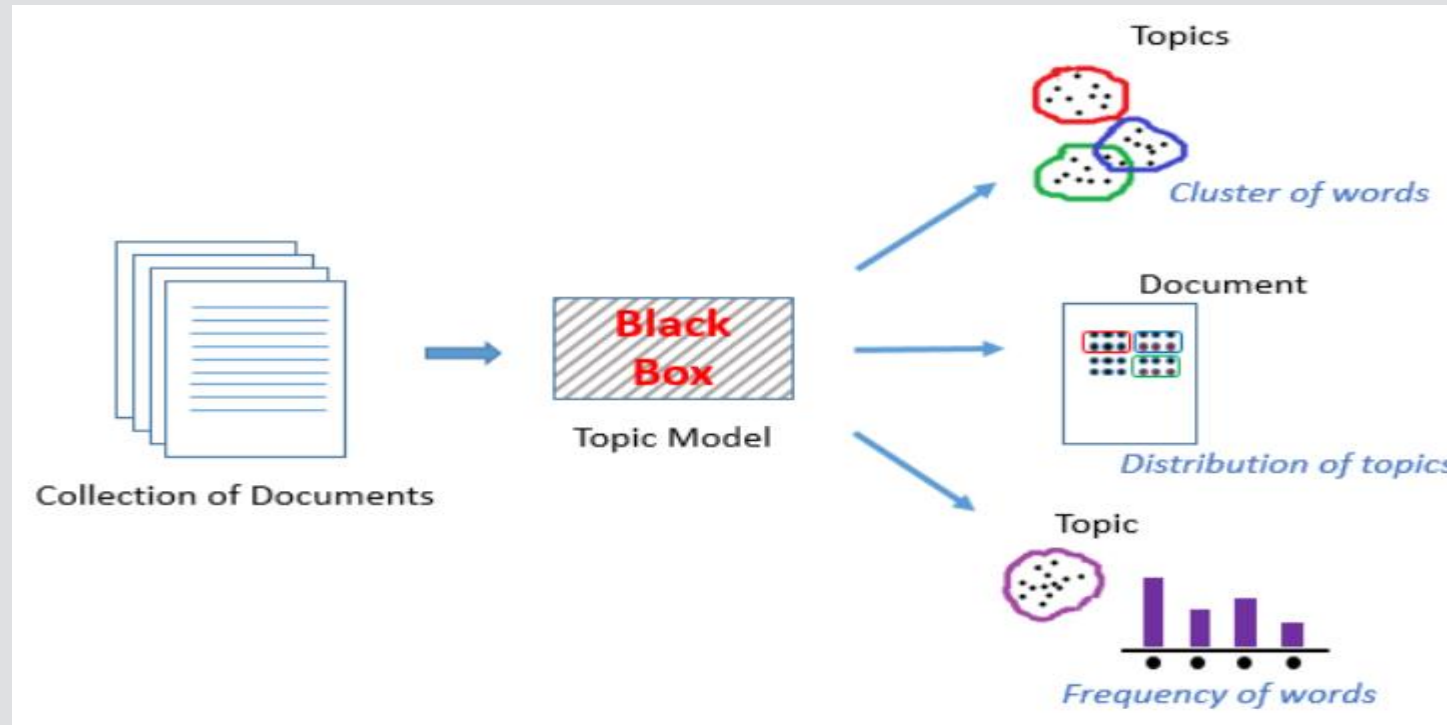


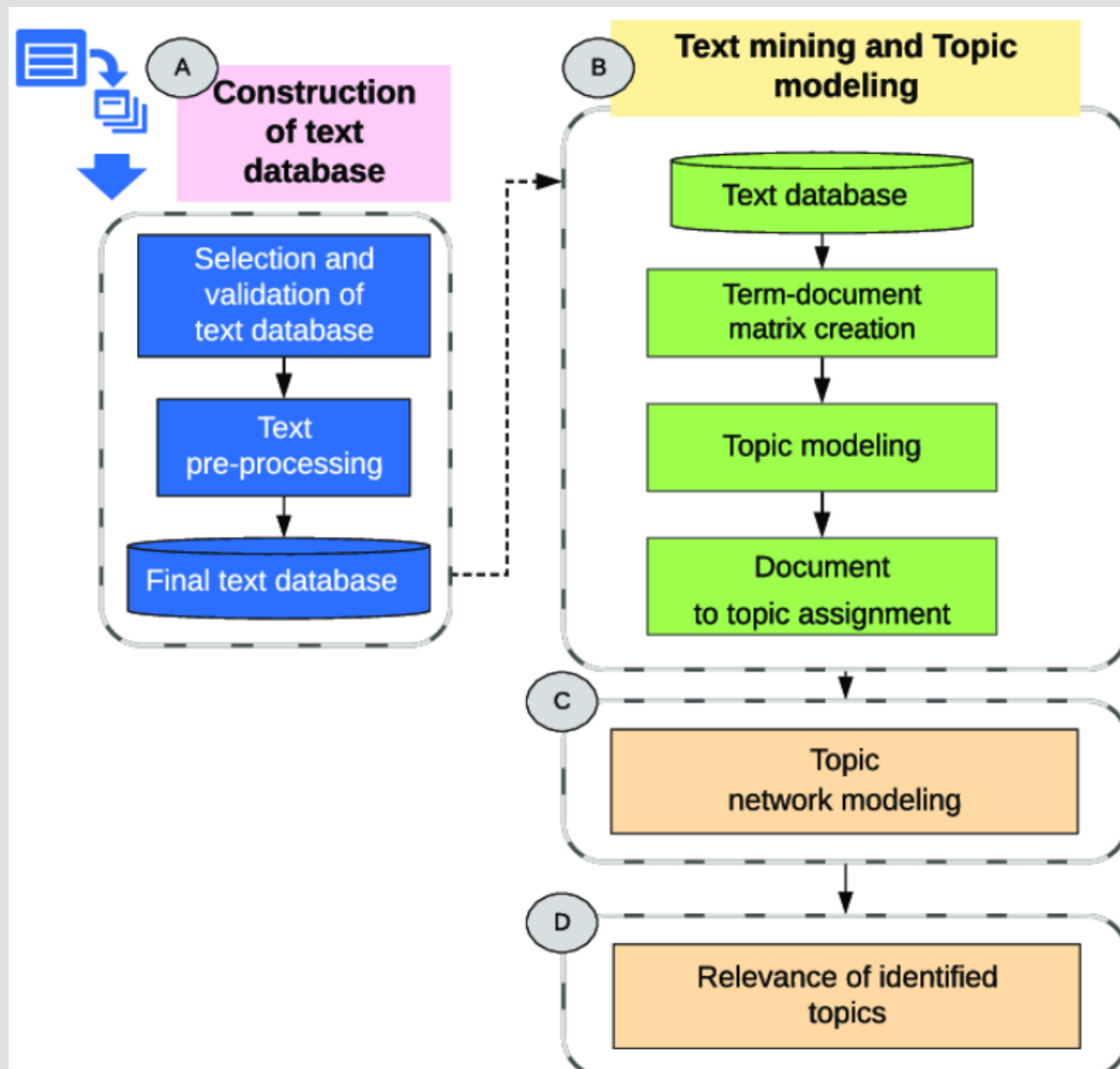
Octubre, 2022

Modelado de Tópicos

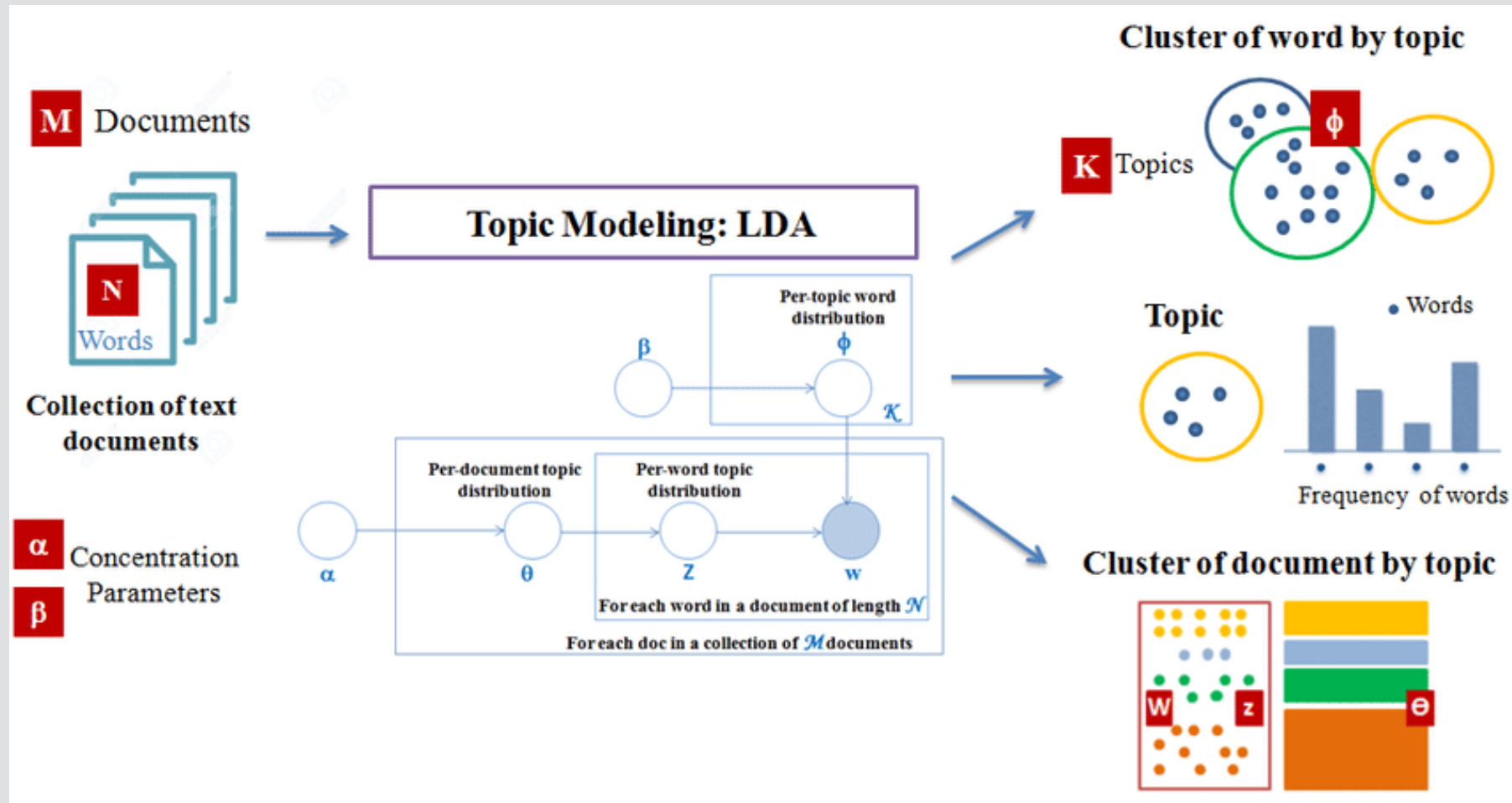
Es una técnica para tratar documentos que no tienen alguna categorización, y asume que cada documento es una mezcla aleatoria de categorías o tópicos

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

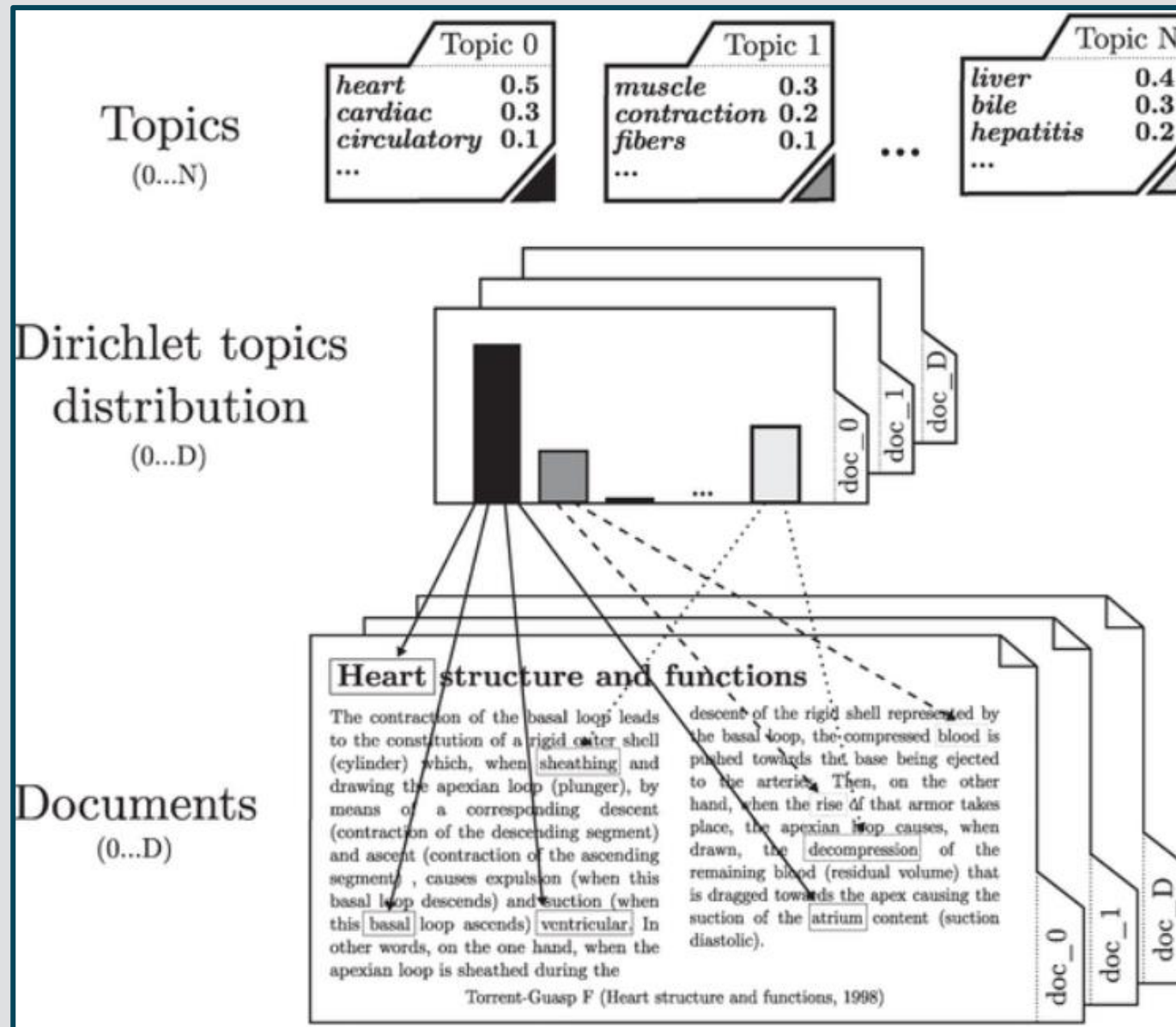




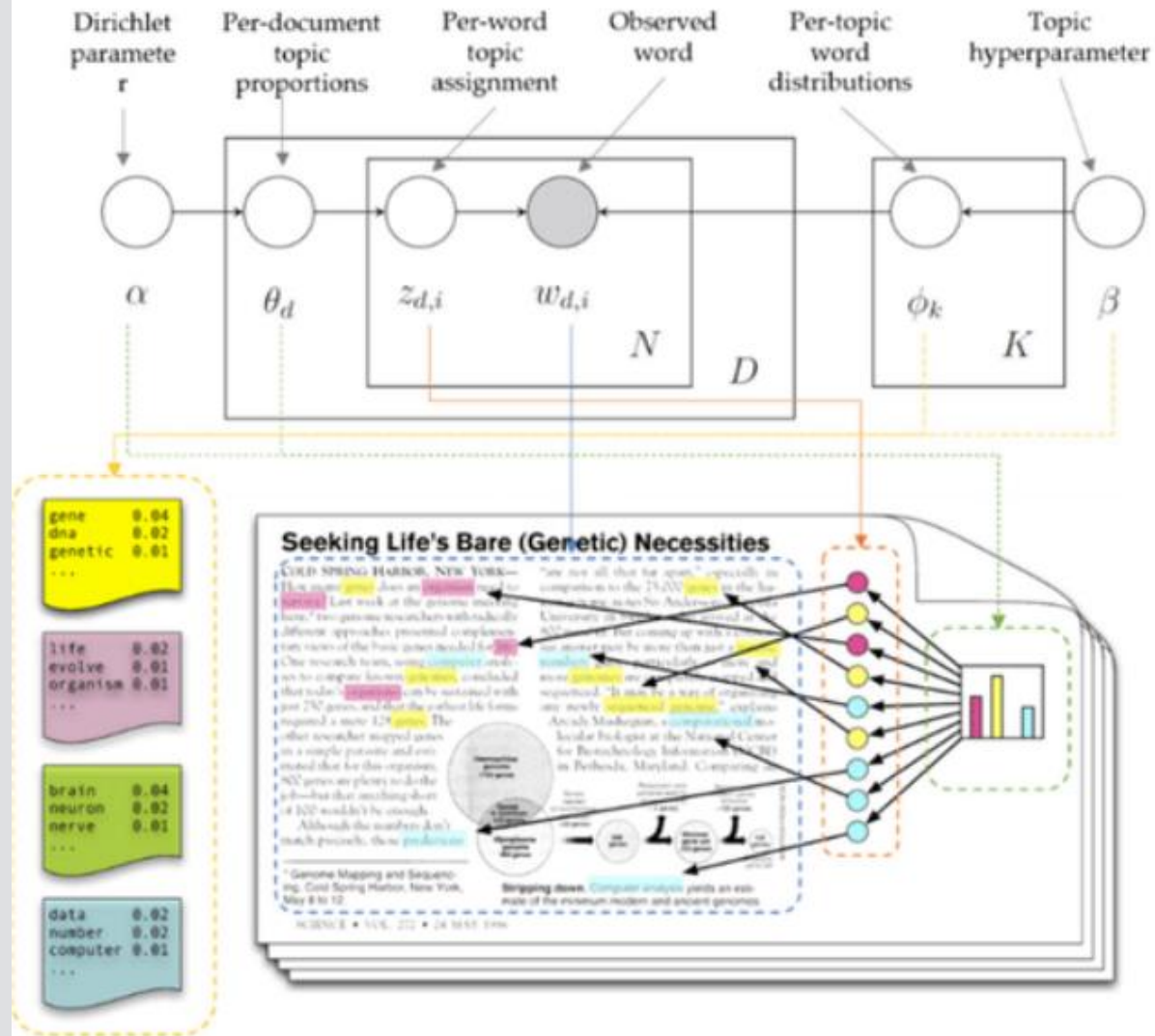
La asignación de Dirichlet latente (LDA) se utiliza como una técnica de modelado de temas que puede clasificar el texto de un documento en un tema (Tópico) en particular. Utiliza la distribución de Dirichlet para encontrar temas o tópicos para cada modelo de documento y palabras para cada modelo de tema.



Tomado de: Amara 2020



(a) LDA document generation process



(b) An illustrative example of LDA document generation process

(c) Two outputs of LDA

(c-1) Per-document topic proportions

	Topic 1	Topic 2
Doc 1	0.20	0.50
Doc 2	0.50	0.02
Doc 3	0.05	0.12
...	...	...
Doc N	0.14	0.25

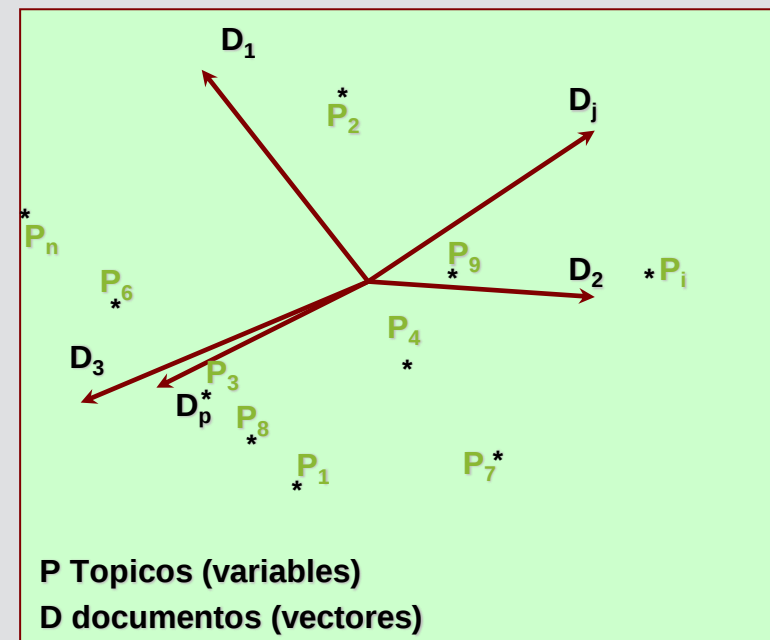
(c-2) Per-topic word distributions

	Topic 1	Topic 2
word 1	0.01	0.05
word 2	0.02	0.02
word 3	0.05	0.12
...	...	...
word N	0.04	0.01

¿Qué es un Biplot?

$$X_{n \times p} = \begin{matrix} & \text{Documentos} \\ \text{Tópicos} & \begin{pmatrix} f_{11} & f_{1p} \\ \vdots & \vdots \\ f_{n1} & f_{np} \end{pmatrix} \end{matrix}$$

Un Biplot (Gabriel, 1971) es una representación gráfica de datos multivariantes. De la misma manera que un diagrama de dispersión muestra la dispersión conjunta de dos variables, un Biplot representa tres o más variables (Gabriel y Odoroff, 1990)



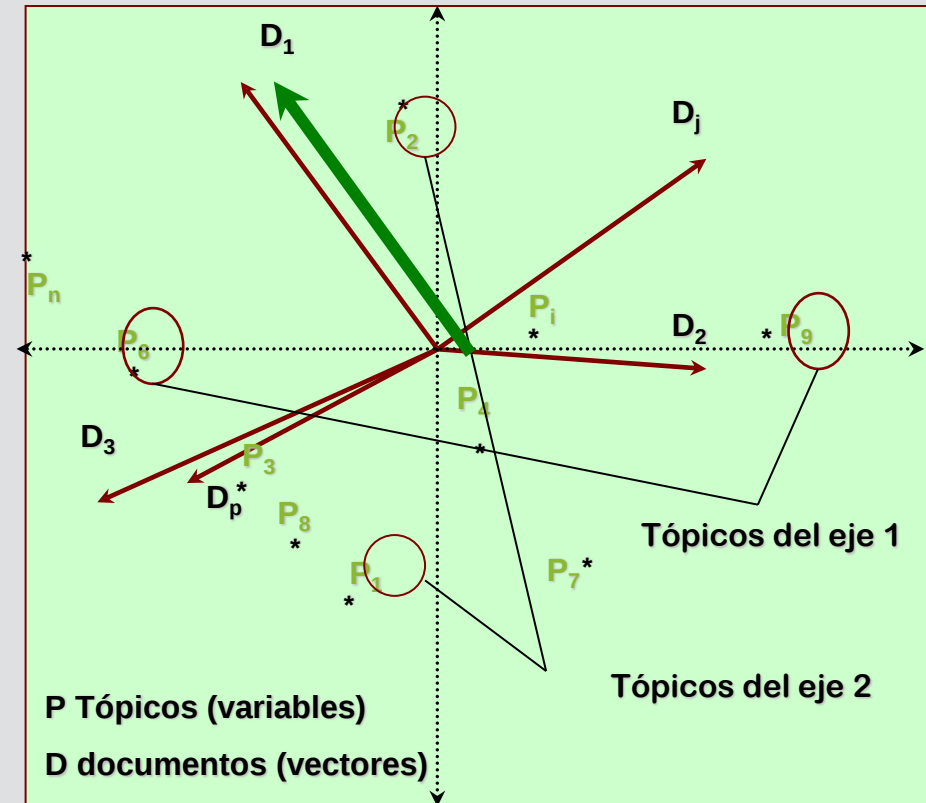
A partir del gráfico se puede conocer:

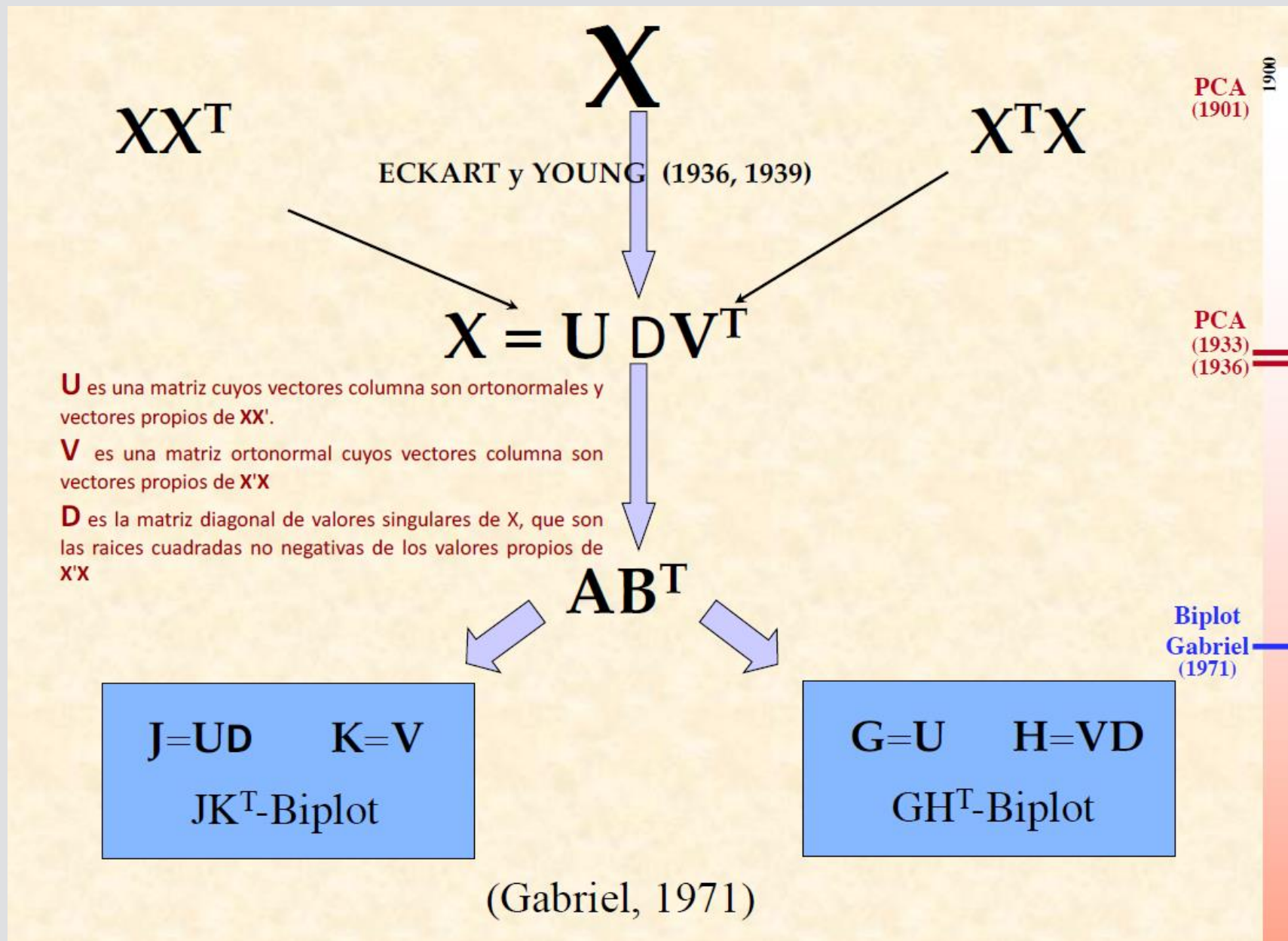
Riqueza léxica

La **variabilidad** de los documentos: observando la **longitud del vector**.

La **similitud** de los documentos: observando el **ángulo entre ellos**.

El **significado latente** de cada eje factorial viene dado por los **tópicos** que tienen altas contribuciones relativas del factor al tópico (tema).





$$X = UDV^T$$

$$J = UD$$

$$H = DV$$




1.- Filas y columnas pueden ser representadas en el mismo sistema de referencia.

a) Puntos fila y puntos columna pueden relacionarse mediante combinaciones lineales simétricas.

b) Ambas nubes presentan la misma dispersión.

2.- La bondad de ajuste es idéntica para filas y para columnas.

REPRESENTACIÓN SIMULTÁNEA	COORDENADAS FILAS	COORDENADAS COLUMNAS	BONDAD AJUSTE PARA FILAS	BONDAD AJUSTE PARA COLUMNAS
GH-BIPLLOT	U	ΣV	$\frac{2}{r}$	$\frac{\lambda_1^2 + \lambda_2^2}{\sum_{\alpha=1}^r \lambda_{\alpha}^2}$
JK-BIPLLOT	$U\Sigma$	V	$\frac{\lambda_1^2 + \lambda_2^2}{\sum_{\alpha=1}^r \lambda_{\alpha}^2}$	$\frac{2}{r}$
HJ-BIPLLOT 	$U \Sigma$	ΣV	$\frac{\lambda_1^2 + \lambda_2^2}{\sum_{\alpha=1}^r \lambda_{\alpha}^2}$	$\frac{\lambda_1^2 + \lambda_2^2}{\sum_{\alpha=1}^r \lambda_{\alpha}^2}$

Octubre, 2022

ORGANIZADORES

datalat.org



AUSPICIANTES

CS&S Code for Science & Society

Chan Zuckerberg Initiative

GORDON AND BETTY
MOORE
FOUNDATION

LDABiplots

Es una herramienta de extracción, análisis y visualización para el análisis exploratorio, permite implementar el modelo probabilístico Latent Dirichlet Allocation (LDA) (Blei, Ng y Jordán, 2003) y generar visualizaciones Biplot (Gabriel K.R, 1971) y HJ-Biplot (Galindo-Villardón P, 1986) , permite optimizar la extracción de datos desde la web, la rutina del modelado LDA y la generación de visualizaciones Biplot de una forma interactiva para usuarios que no estén adaptados al uso de R.

Cran de paquete

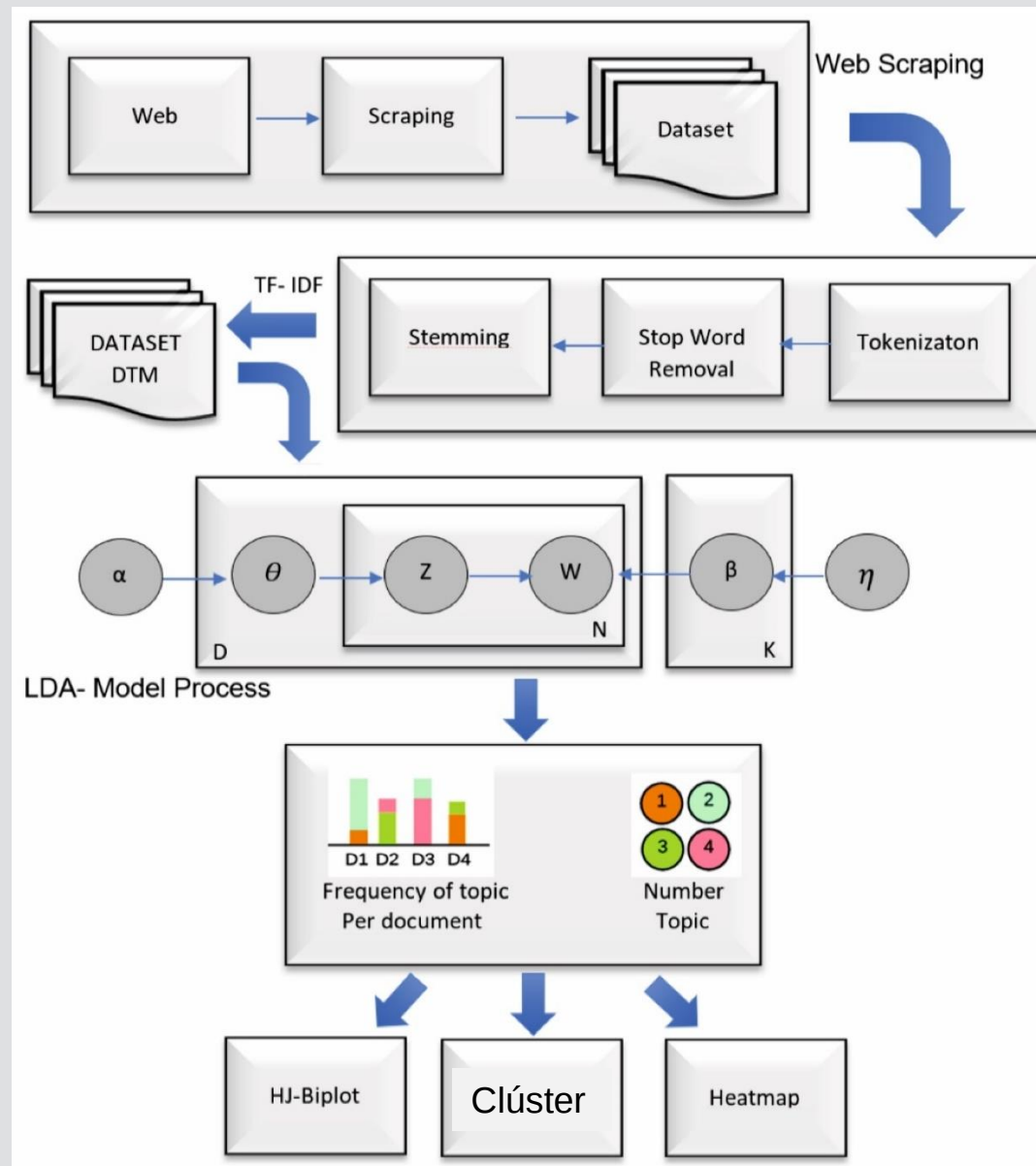
<https://cran.r-project.org/web/packages/LDABiplots/index.html>

Tutorial de paquete

https://cran.r-project.org/web/packages/LDABiplots/vignettes/Tutorial_LDABiplots_Spanish.html

Video Tutorial de paquete

<https://youtu.be/nsOUE9NM9is>



R version 4.2.0 (2022-04-22 ucrt)

```
install.packages("LDABiplots")  
library(LDABiplots)  
runLDABiplots()
```

R: Biplot Graphical Interface for LDA Models ▾ Find in Topic

Biplot Graphical Interface for LDA Models

Documentation for package 'LDABiplots' version 0.1.2

- [DESCRIPTION file.](#)
- [User guides, package vignettes and other documentation.](#)

Help Pages

dtmcorr	Pearson Correlation for Sparse Matrices
dtmremovetfidf	Remove terms from a Document-Term-Matrix and documents with no terms based on the term frequency inverse document frequency
dtmtfidf	Term Frequency - Inverse Document Frequency calculation
GHBiplot	GHBiplot
HJBiplot	HJBiplot
JKBiplot	JKBiplot
Plot_Biplot	Plotting Biplot
runLDABiplots	Shiny UI for LDABiplots package

Octubre, 2022

ORGANIZADORES

datalat.org



AUSPICIANTES



LDABiplots

About

Import or Load Data

Document Term Matrix Visualizations

Inference (Number of topic)

LDA and Biplot

LDABiplots

LDABiplots

LDABiplots

Es una herramienta de extracción, análisis y visualización para el análisis exploratorio de noticias publicadas en la web por periódicos digitales, mediante la extracción de datos de la web (Bradley et al. 2019), permite la implementación del modelo probabilístico Latent Dirichlet Allocation (LDA) (Blei, Ng y Jordán, 2003) y generar visualizaciones Biplot (Gabriel K.R, 1971) y HJ-Biplot (Galindo-Villardón P, 1986) de los temas principales de los titulares de las noticias publicadas en la web. LDABiplots permite optimizar la extracción de

Ahora nos vamos a la práctica

Referencias:

Blei, D., Carin, L., & Dunson, D. (2010). Probabilistic topic models. *IEEE signal processing magazine*, 27(6), 55-65.

Villardón- Galindo, M. P. (1986). Una alternativa de representación simultánea: HJ-Biplot. *Qüestiió: quaderns d'estadística i investigació operativa*, 13-23.

Gabriel, K. R. (1980). *Biplot display of multivariate matrices for inspection of data and diagnosis*. ROCHESTER UNIV NY.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan), 993-1022.

Pilacuan-Bonete, L., Galindo-Villardón, P., & Delgado-Álvarez, F. (2022). HJ-Biplot as a Tool to Give an Extra Analytical Boost for the Latent Dirichlet Assignment (LDA) Model: With an Application to Digital News Analysis about COVID-19. *Mathematics*, 10(14), 2529.

Marin-Zulaima. C. O. (2006). *Contribuciones al análisis de datos textuales* (Doctoral dissertation, Universidad de Salamanca).

CO-ORGANIZADORES



ALIADOS DE DIFUSIÓN



ORGANIZADORES

datalat.org



AUSPICIANTES

CS&S

Code for Science & Society

Chan Zuckerberg Initiative

GORDON AND BETTY MOORE FOUNDATION

Octubre, 2022