

# Minería de textos: un enfoque distinto de la lengua



# HEY!!!

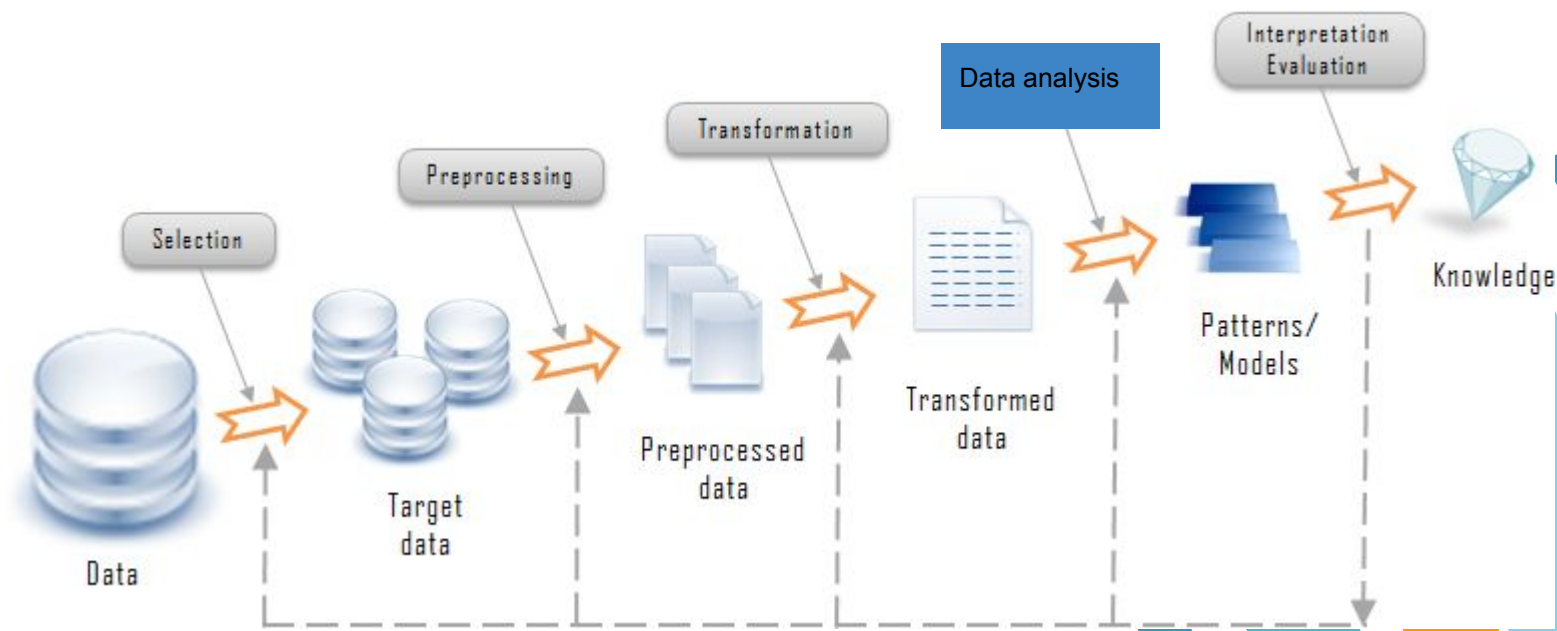
## Soy Héctor Murrieta

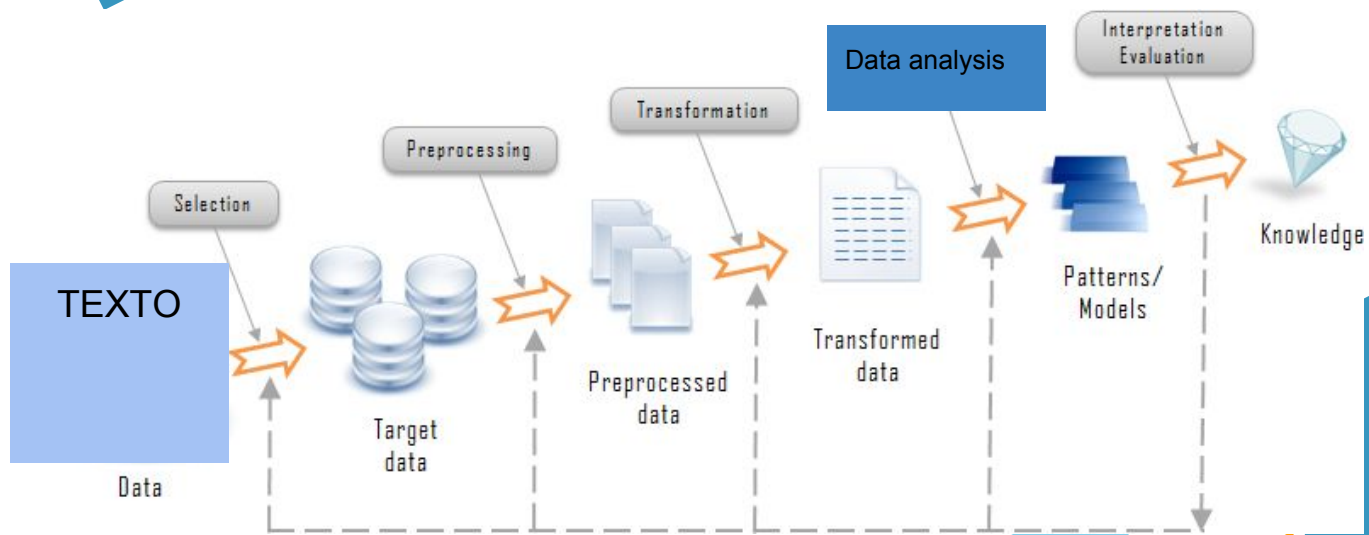
Y digo que sé deep learning

[hector@mariachi.io](mailto:hector@mariachi.io)



# En general ¿qué es minar datos?







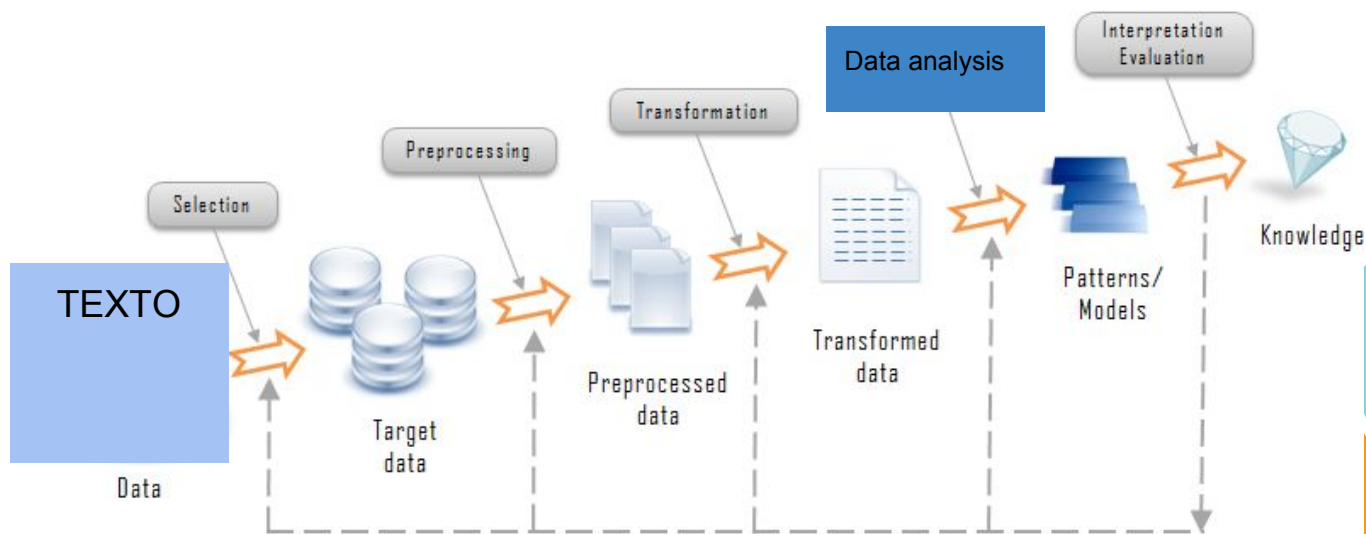
**Para qué sirve obtener éste conocimiento:**

**-crear productos**

**-entender grandes cantidades de texto**



# ¿Qué producto creen que siga éste patrón?





# Google

Google Search

I'm Feeling Lucky

*Behind today's changes...*



**¿Qué preguntas no responde google\*?**



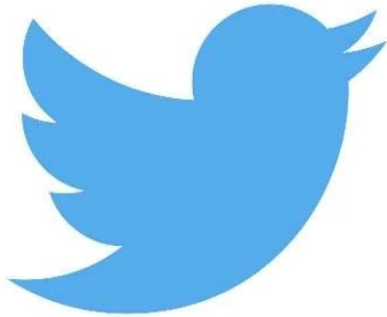




Springer



IEEE





# **Técnicas comunes**



# Keywords:

- más fácil
- operación lógica
- mas simple

Your product or service

asian food near me Get ideas Modify search

Ad group ideas Keyword ideas Columns Download Add all (801)

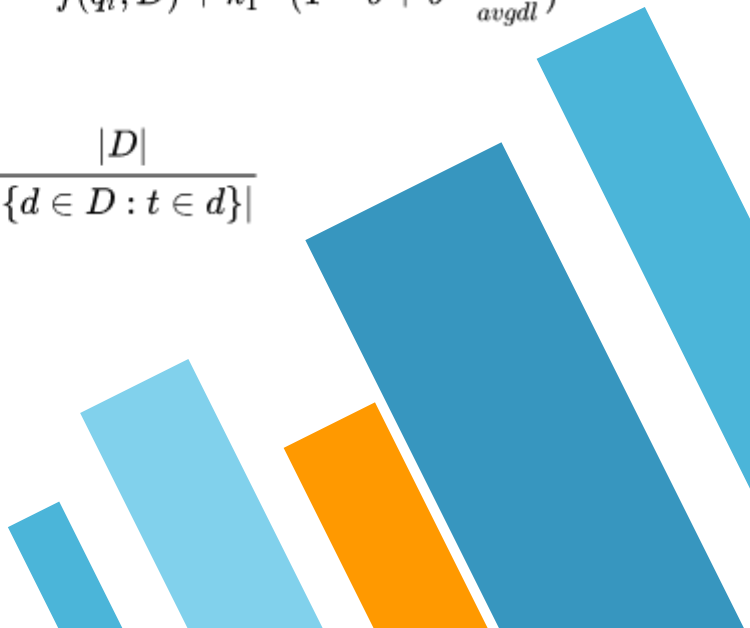
Keyword (by relevance)	Avg. monthly searches	Competition	Suggested bid	Ad impr.	Add to plan
chinese takeaway delivery near me	10	High	UAH99.90		»
fast food delivery near me	4,400	Medium	UAH69.57		»
who delivers food near me	2,400	Medium	UAH58.59		»
local chinese takeaway delivery	10	Medium	UAH48.38		»
food deliveries near me	2,400	Medium	UAH72.30		»
nearby restaurants that deliver	1,300	Medium	UAH63.78		»
find chinese food delivery	20	Medium	UAH53.03		»



## Vectorización:

- Mejores resultados
- Más compleja
- muy popular últimamente

$$\text{score}(D, Q) = \sum_{i=1}^n \text{IDF}(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot (1 - b + b \cdot \frac{|D|}{\text{avgdl}})}$$

$$\text{idf}(t, D) = \log \frac{|D|}{|\{d \in D : t \in d\}|}$$




# **Keywords vs Vectorización**



**Apple** CEO **Tim Cook** Introduces 2 New, Larger iPhones, Smart Watch At **Cupertino** **Flint Center** Event

Person

Organisation

Location

*Figure 1: An example of NER application on an example text*



Search worldwide, life-sciences literature

 Search

[Advanced Search](#)

E.g. "breast cancer" HER2 Smith J



### Search more than abstracts

- **Abstracts** (33.3 million, including 28.2 million from PubMed)
- **Full text articles** (4.6 million)
- **Patents** (4.2 million)
- **Agricola records** (675,698)
- **NHS clinical guidelines** (860)

[About Europe PMC](#) ⓘ



### Link to public databases

Explore protein, gene, species and disease records **directly from articles**:

- UniProt
- Protein Data Bank (PDBe)
- European Nucleotide Archive (ENA)
- Wikipedia and other lay summaries

[Learn how we use text-mining](#) ⓘ



### Get credit for your work

ORCID is a unique identifier for researchers which distinguishes you from every other researcher, and makes it easier to find your work.

Use our claiming tool to link your Europe PMC articles to your ORCID

[Link articles to your ORCID](#) ⓘ

#### About

[About Europe PMC](#)  
[Funders](#)  
[Joining Europe PMC](#)  
[Governance](#)  
[Roadmap](#)

#### Tools

[Tools overview](#)  
[ORCID article claiming](#)  
[Journal list](#)  
[Grant finder](#)  
[External links service](#)

#### Developers

[Developer resources](#)  
[Articles RESTful API](#)  
[Grants RESTful API](#)  
[SOAP web service](#)  
[Annotations API](#)

#### Help

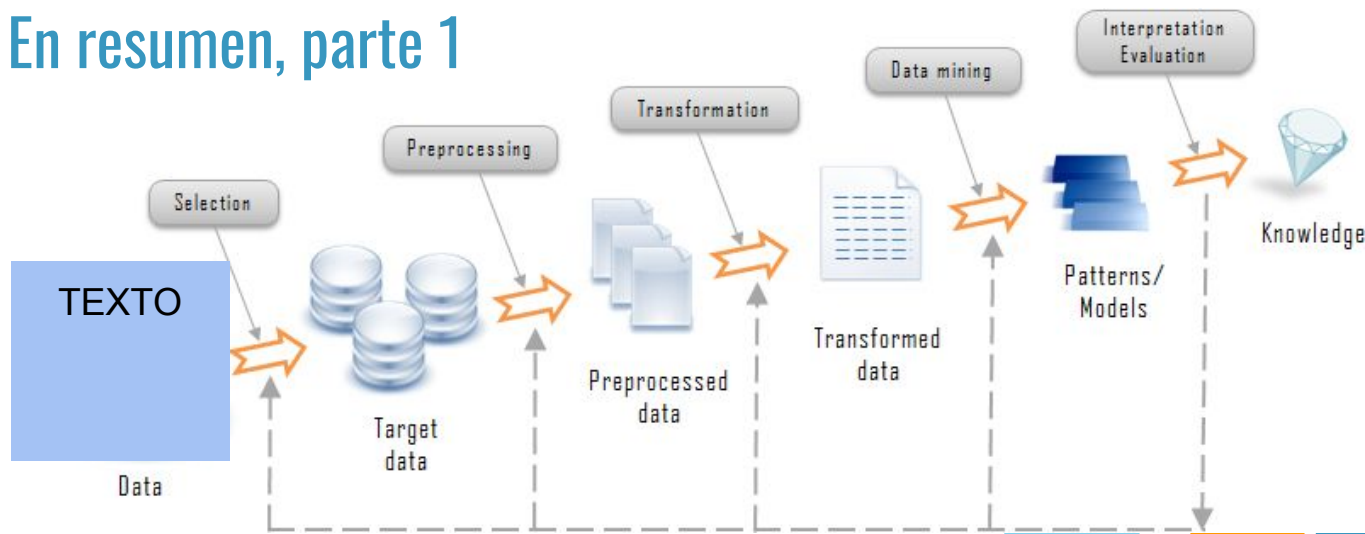
[Help using Europe PMC](#)  
[Contact us](#)

#### Contact us



[Helpdesk](#)  
[Feedback](#)  
[Twitter](#)  
[Blog](#)

 Feedback

## En resumen, parte 1







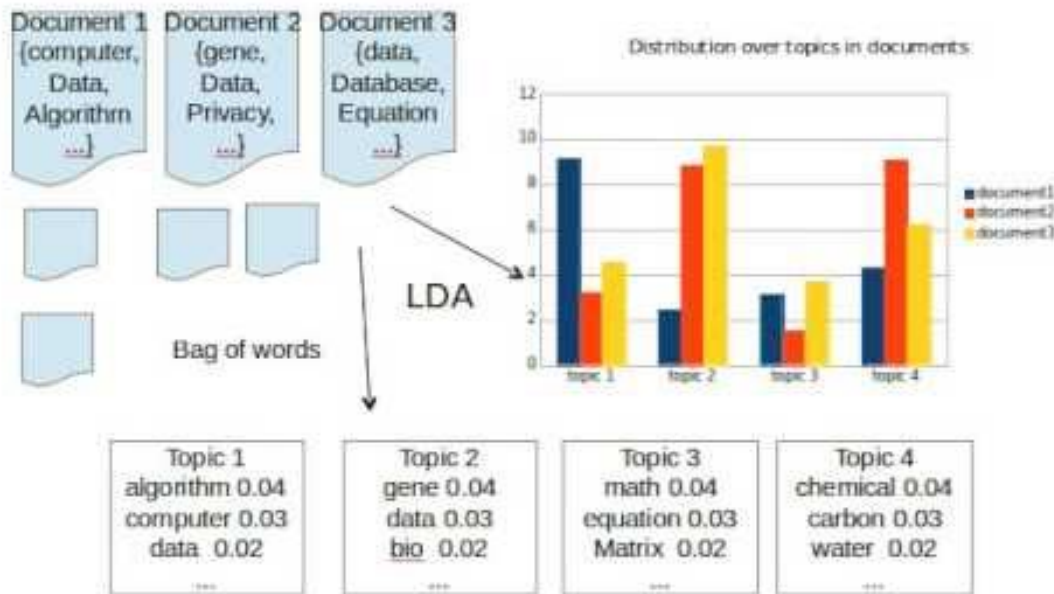
## Parte 2

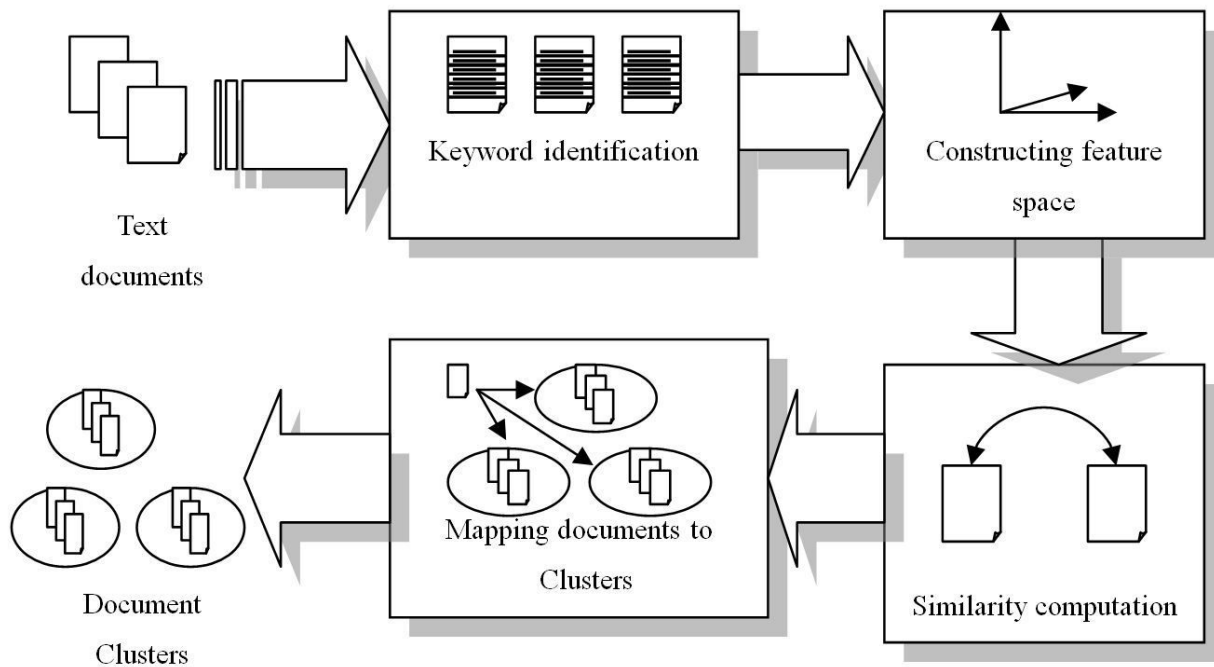
# Minar texto como análisis del discurso



¿Qué es el análisis del discurso?

# LDA









Diferentes herramientas, otro  
tipo de aplicaciones





# ¡Gracias!

## ¿Alguna pregunta?

¿Tienes tiempo para hablar de chatbots?

- » Bots LATAM
- » [hector@mariachi.io](mailto:hector@mariachi.io)

