Doctorado en
Ciencias Informáticas

Doctorado en Tecnologías
Informáticas Avanzadas

Generación automática inteligente de resúmenes de textos con técnicas de Soft Computing

Tesis doctoral realizada en cotutela

Alumno:

Lic. Augusto Villa Monte

Directores:

Dra. Laura Lanzarini - UNLP
Dr. José A. Olivas - UCLM

~ Marzo de 2019 ~

Agenda

01 Motivación y objetivos

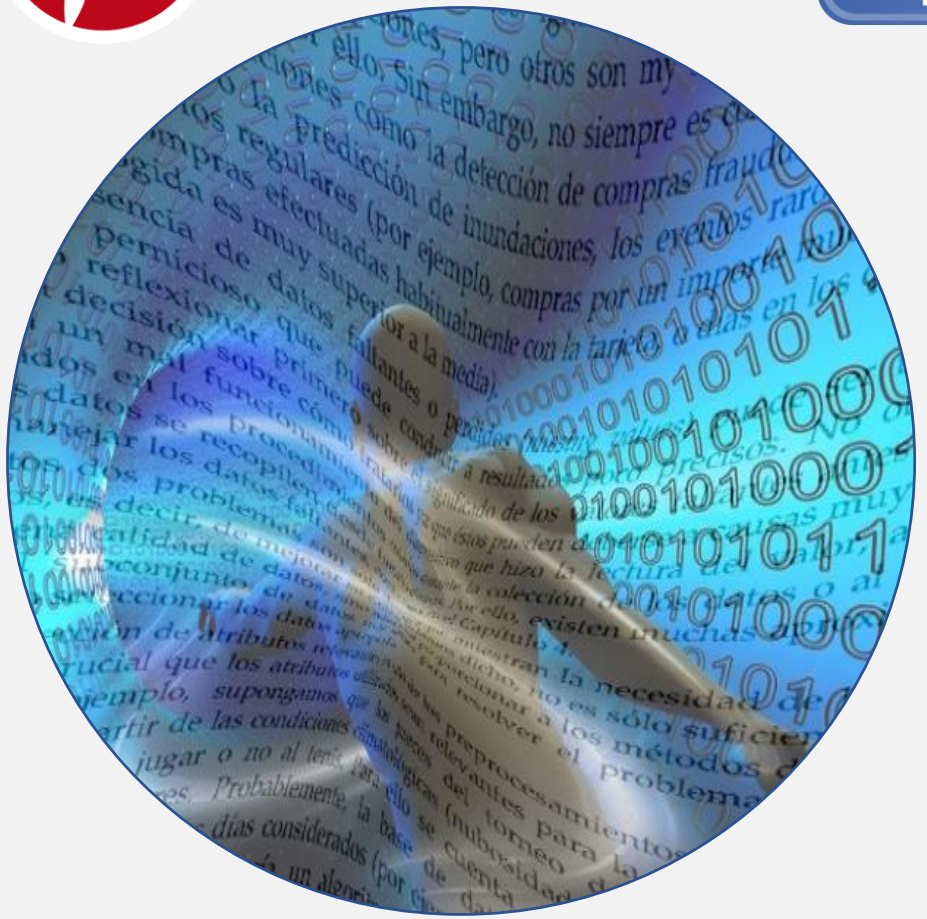
02 Procesamiento de texto y obtención de resúmenes automáticos

03 Resumen utilizando una técnica de optimización mediante cúmulo de partículas

04 Resumen mediante grafos causales y con componentes temporales

05 Conclusiones y trabajo futuro

Motivación



Motivación

El ser humano almacena el conocimiento en documentos

Y consume continuamente información textual

https://cdn-images-1.medium.com/max/2600/1*wvjR1rGD85div2WufLtgHA.jpeg

Motivación

Sería ideal que el ser humano pudiera recordar absolutamente todo lo que lee

Capta la información esencial para mantenerla en su memoria

https://cdn-images-1.medium.com/max/2600/1*TWegc2eAXH_PSJyQRKagXA.png

Motivación

Resumir texto pretende reducir los problemas generados por el crecimiento desmedido de información textual

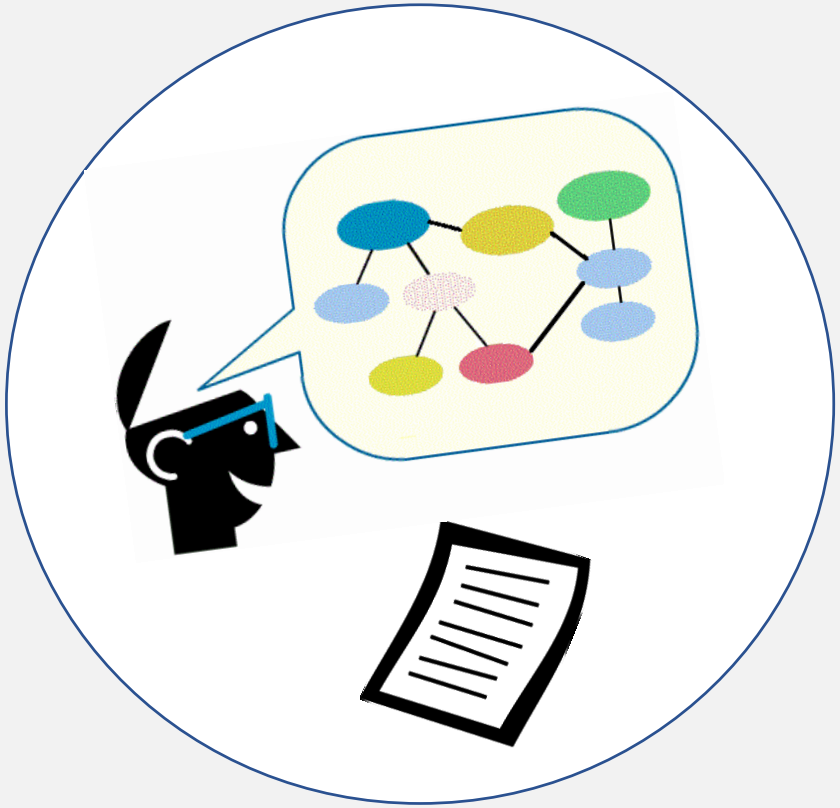
https://blog.frase.io/wp-content/uploads/2018/11/content-[illegible]ting-ai.jpg

Objetivos

Contribuir al área conformada por el PLN y la MT con **dos estrategias** capaces de:

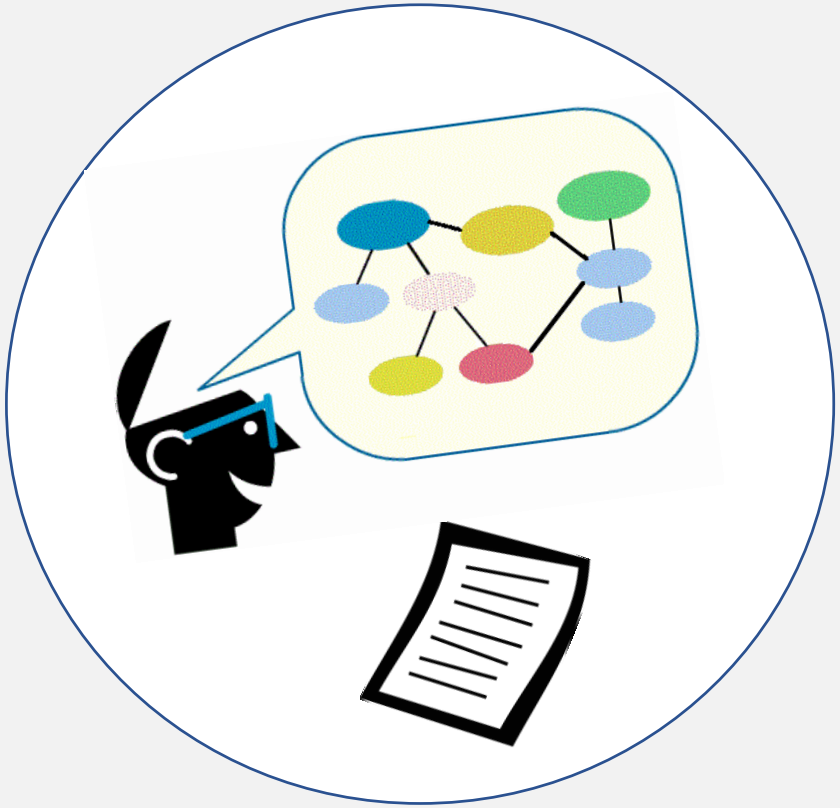
- identificar a partir de un conjunto de documentos lo **relevante**
- y construir con eso un **resumen en forma automática**.

Objetivos



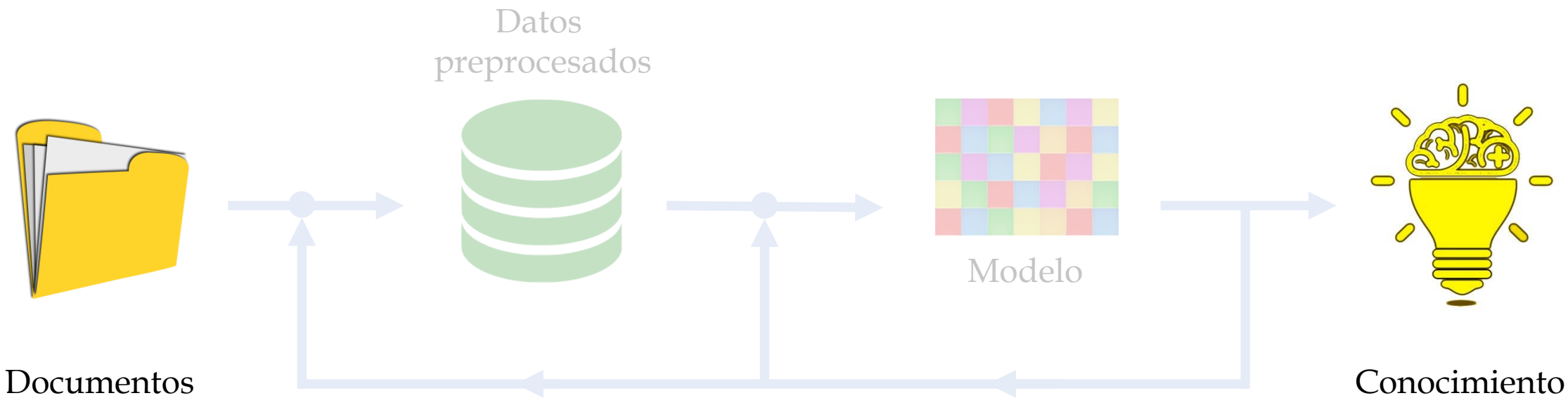
Representación vectorial (métricas)

Objetivos

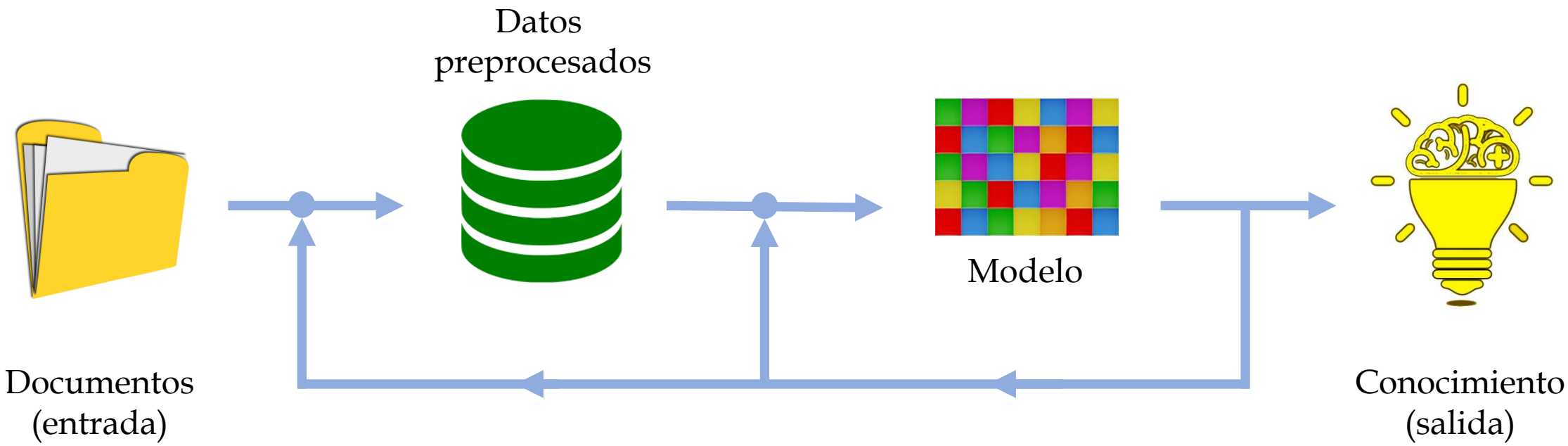


Representación tipo grafo (patrones textuales)

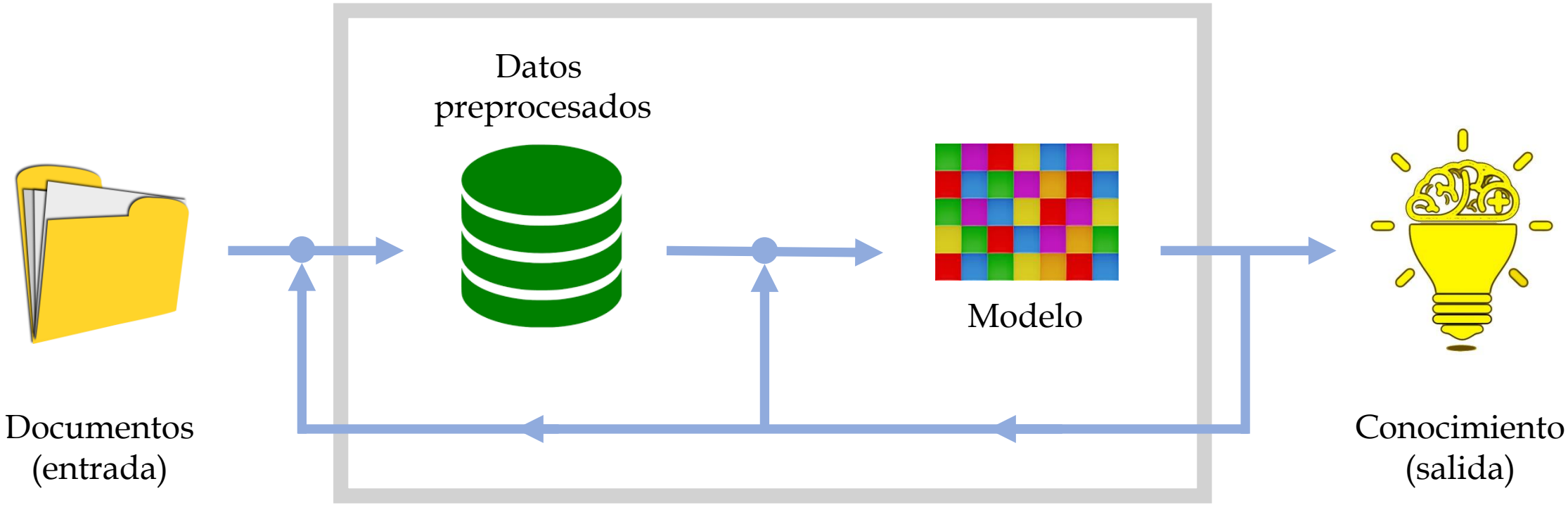
Proceso de KDT



Proceso de KDT



Proceso de KDT



Preprocesamiento del texto

Segmentación y tokenización

Eliminación del ruido

Normalización y filtrado

Preprocesamiento del texto

Etiquetas HTML o XML, encabezados o pies de páginas, publicidad, etc.

Segmentación y tokenización

Eliminación del ruido

Normalización y filtrado

Preprocesamiento del texto

Particionar el texto en componentes significativos

Segmentación y tokenización

Eliminación del ruido

Normalización y filtrado

Preprocesamiento del texto

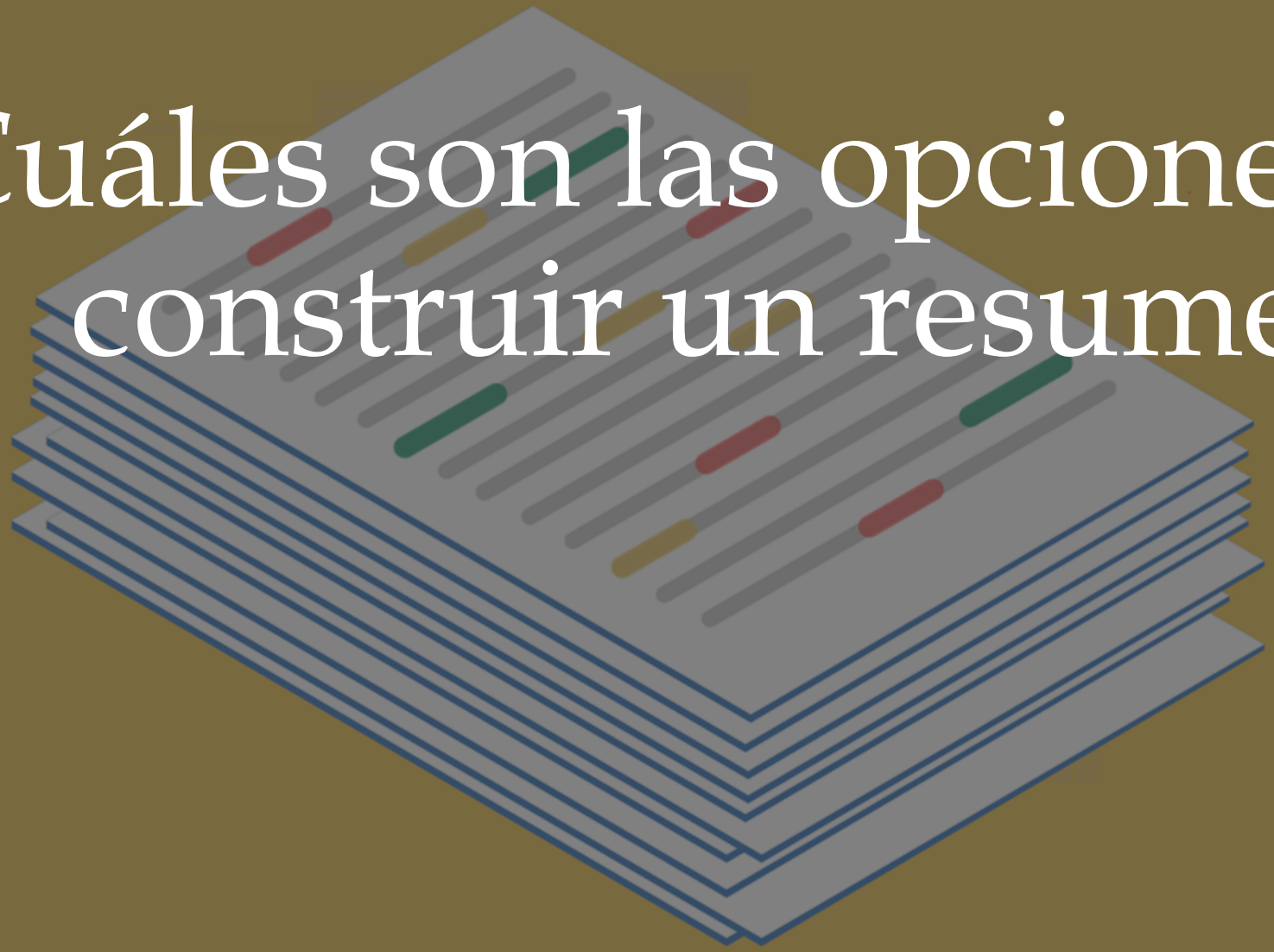
Segmentación y tokenización

Reducción del vocabulario: stemming, lematización, stopwords, tagging

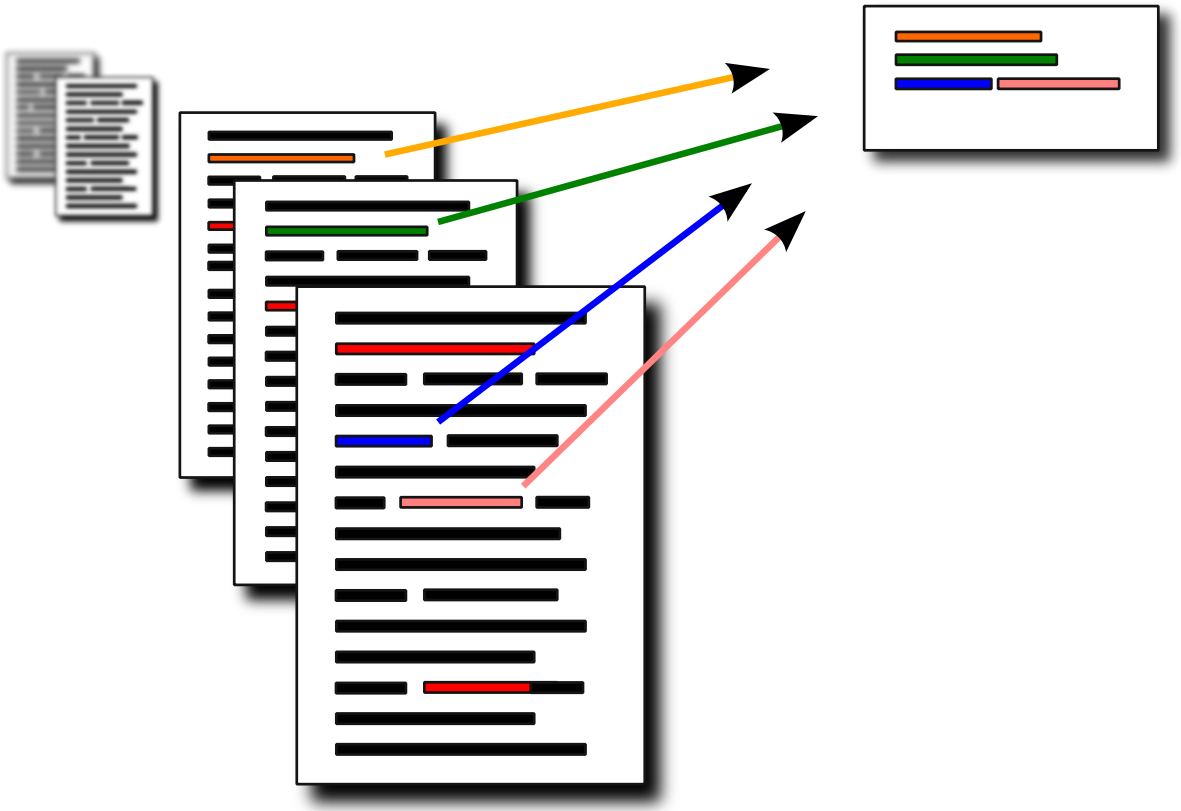
Eliminación del ruido

Normalización y filtrado

¿Cuáles son las opciones para construir un resumen?

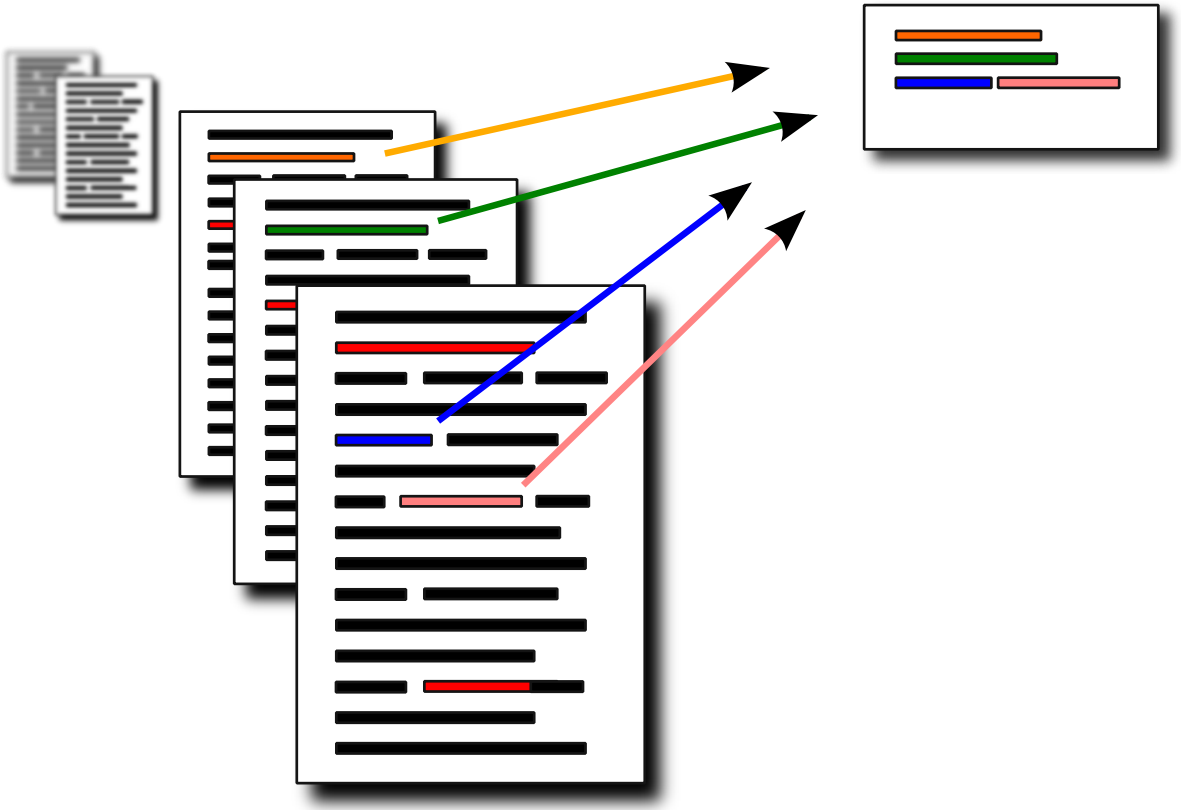


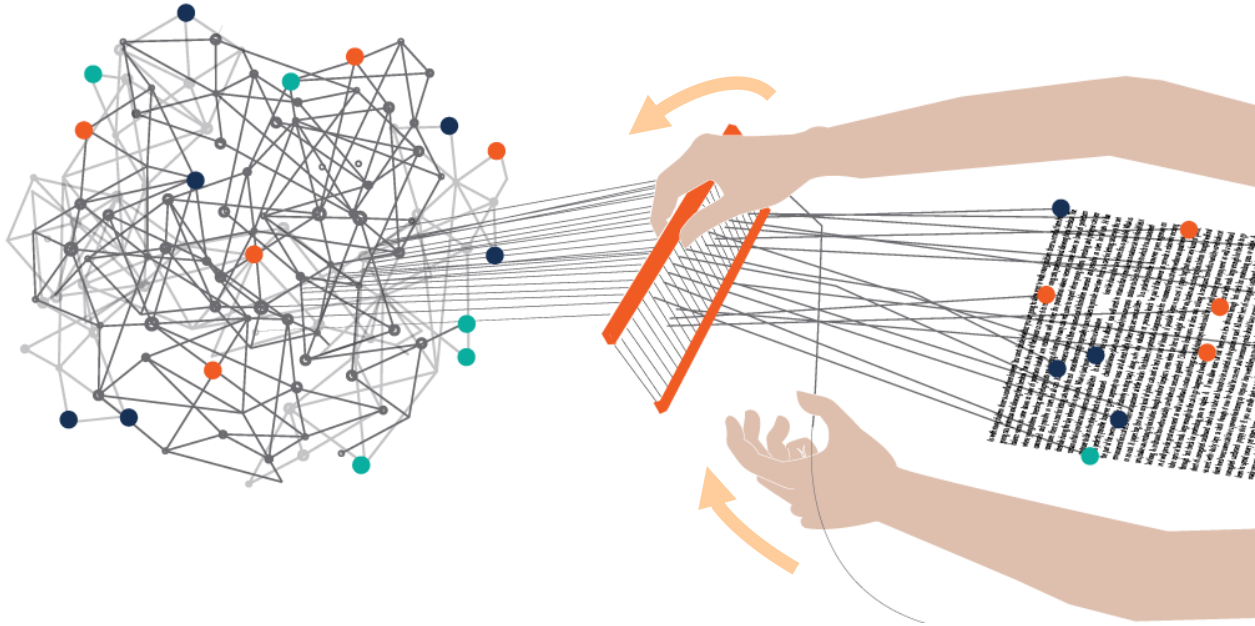
Enfoque extractivo



http://mogren.one/graphics/illustrations/mogren_summarization.svg

Enfoque extractivo vs. abstractivo



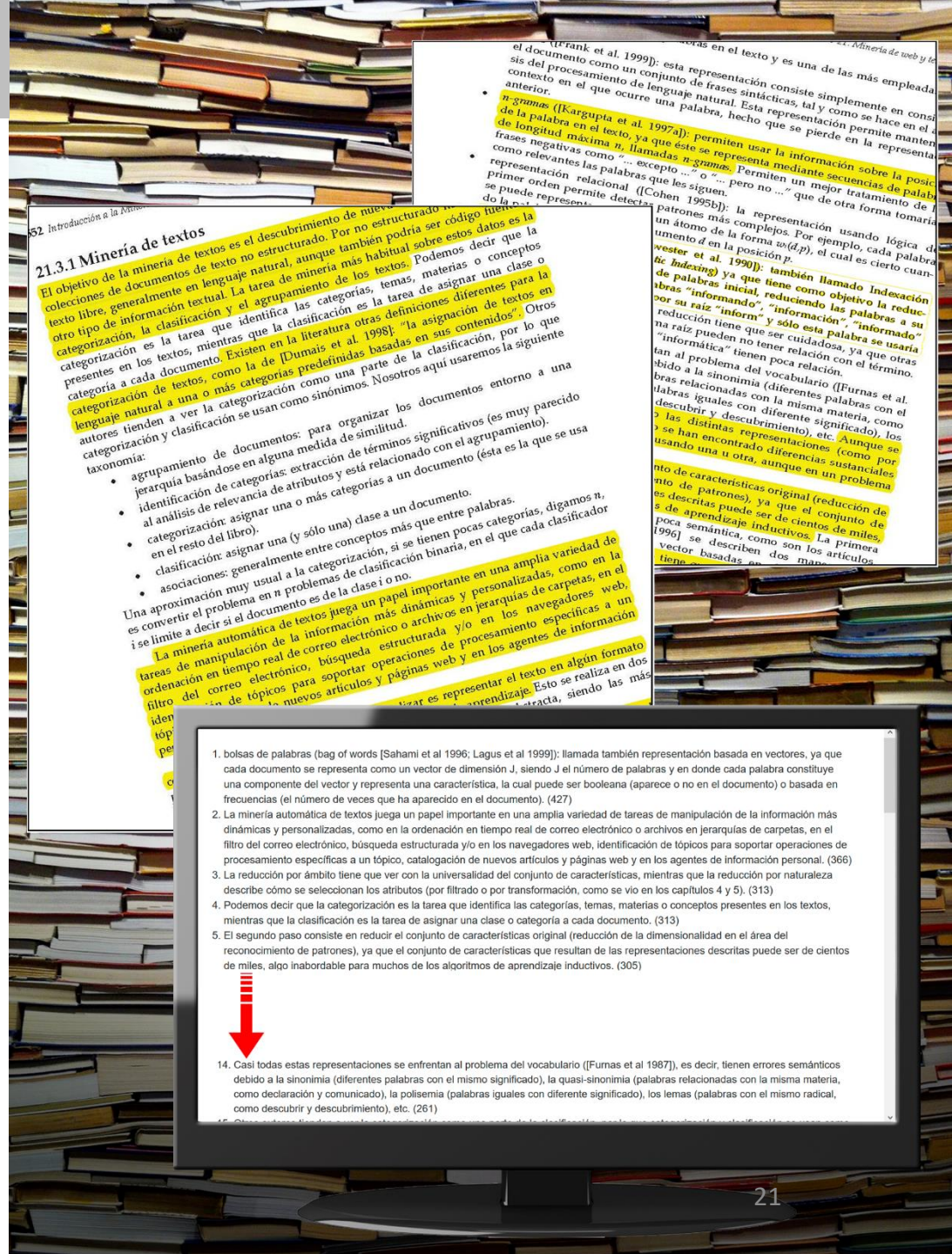


http://mogren.one/graphics/illustrations/mogren_summarization.svg

https://www.ontotext.com/wp-content/uploads/2017/05/Weaving-Data-into-Text.png

Resumen extractivo

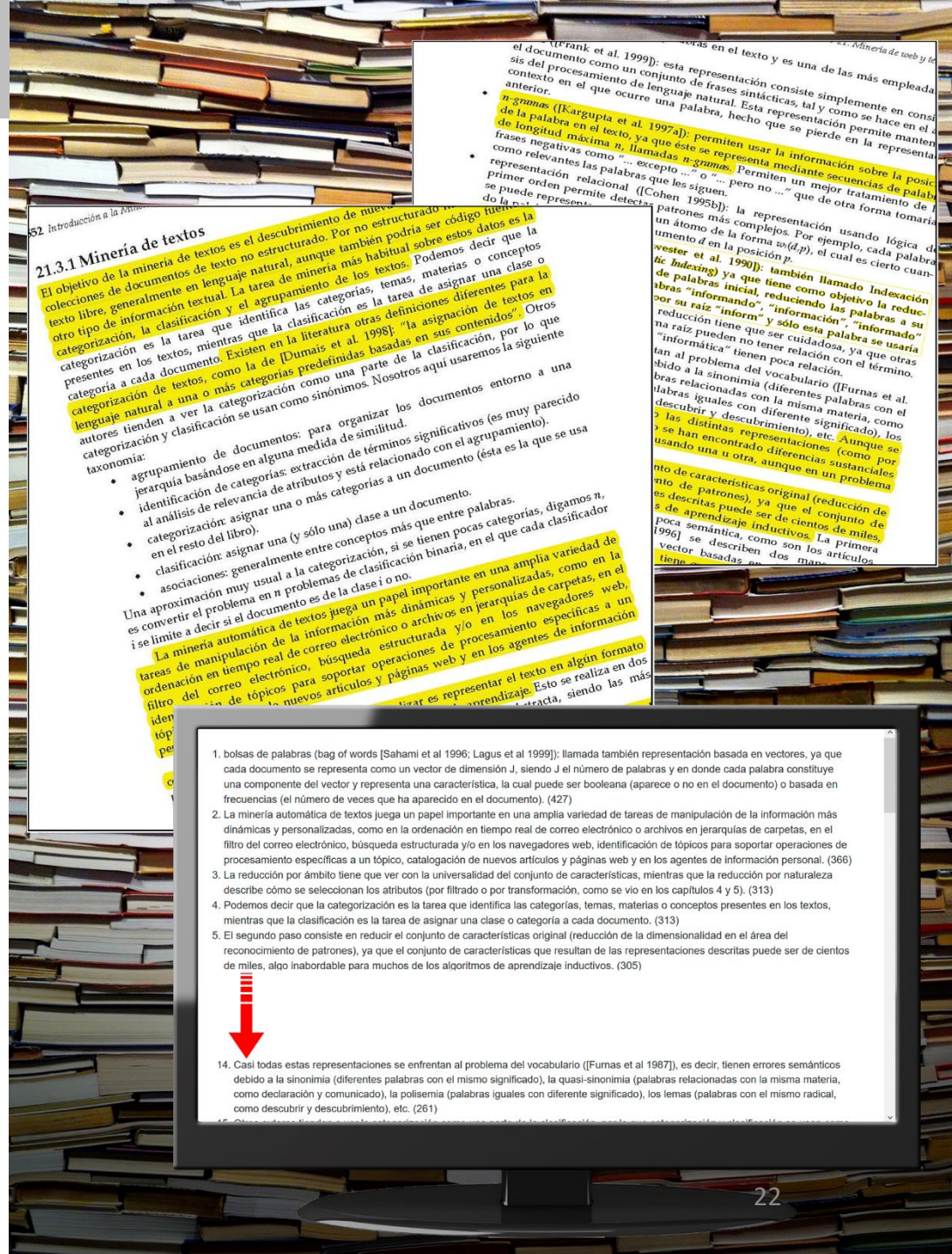
1. Posicion
2. Longitud
3. Keywords
4. Frecuencia
5. Titulos
6. Cobertura



Resumen extractivo

21.3.1 Minería de textos

El objetivo de la minería de textos es el descubrimiento de nueva información a partir de colecciones de documentos de texto no estructurado. Por no estructurado nos referimos a texto libre, generalmente en lenguaje natural, aunque también podría ser código fuente u otro tipo de información textual. La tarea de minería más habitual sobre estos datos es la categorización, la clasificación y el agrupamiento de los textos. Podemos decir que la categorización es la tarea que identifica las categorías, temas, materias o conceptos presentes en los textos, mientras que la clasificación es la tarea de asignar una clase o categoría a cada documento.

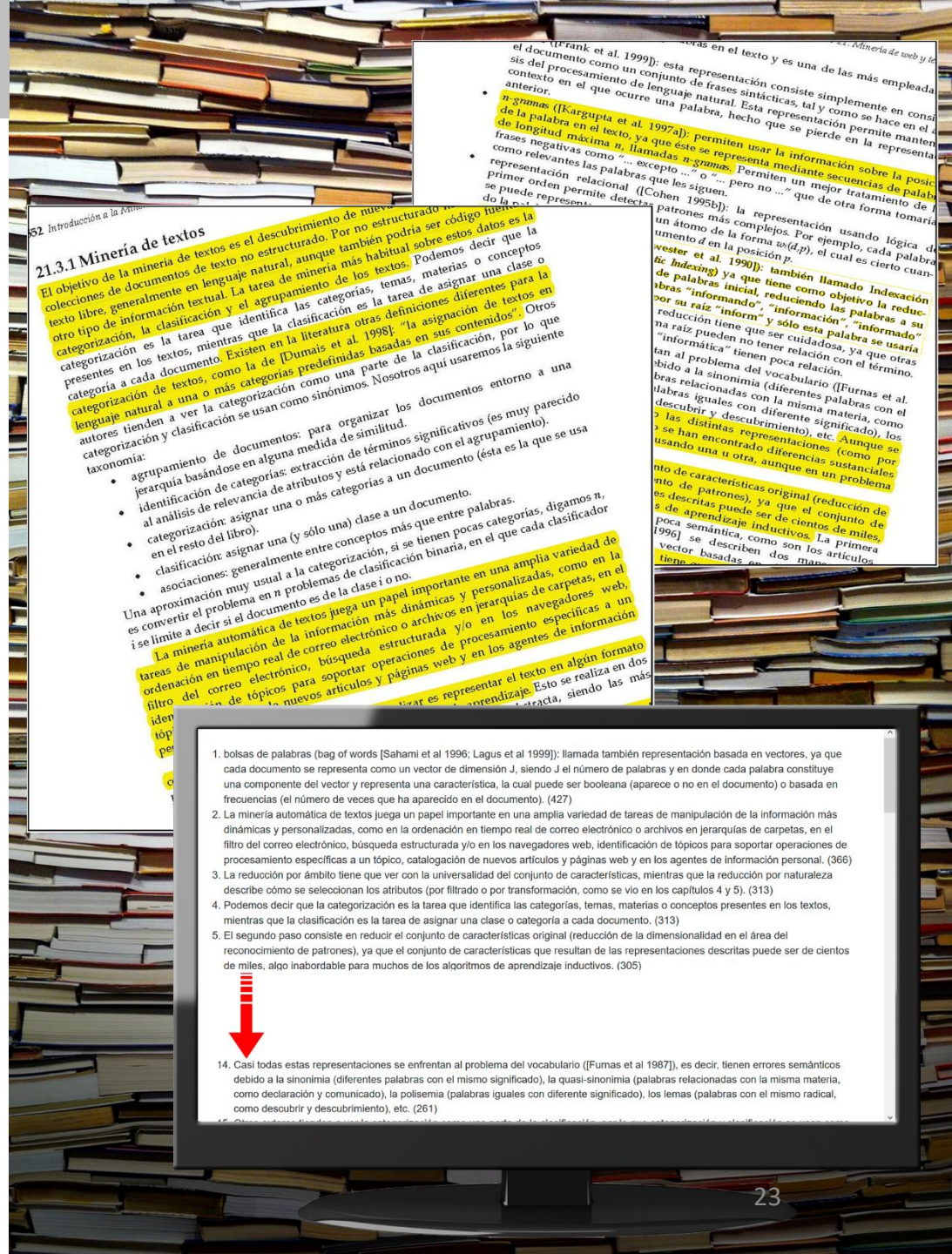


Resumen extractivo

POS_F

21.3.1 Minería de textos

El objetivo de la minería de textos es el descubrimiento de nueva información a partir de colecciones de documentos de texto no estructurado. Por no estructurado nos referimos a texto libre, generalmente en lenguaje natural, aunque también podría ser código fuente u otro tipo de información textual. La tarea de minería más habitual sobre estos datos es la categorización, la clasificación y el agrupamiento de los textos. Podemos decir que la categorización es la tarea que identifica las categorías, temas, materias o conceptos presentes en los textos, mientras que la clasificación es la tarea de asignar una clase o categoría a cada documento.

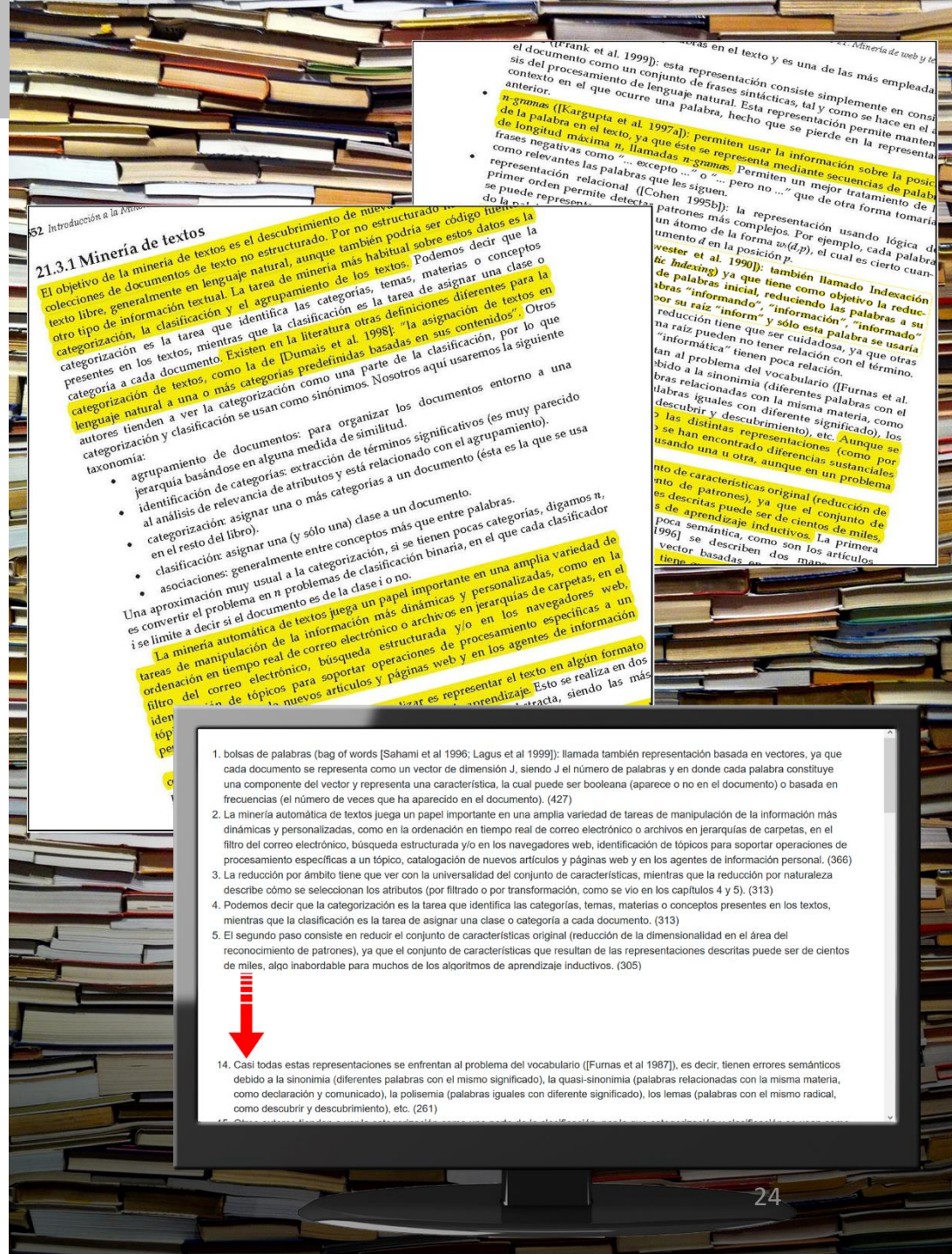


Resumen extractivo

21.3.1 Minería de textos

El objetivo de la minería de textos es el descubrimiento de nueva información a partir de colecciones de documentos de texto no estructurado. Por no estructurado nos referimos a texto libre, generalmente en lenguaje natural, aunque también podría ser código fuente u otro tipo de información textual. La tarea de minería más habitual sobre estos datos es la categorización, la clasificación y el agrupamiento de los textos. Podemos decir que la categorización es la tarea que identifica las categorías, temas, materias o conceptos presentes en los textos, mientras que la clasificación es la tarea de asignar una clase o categoría a cada documento.

POS_L

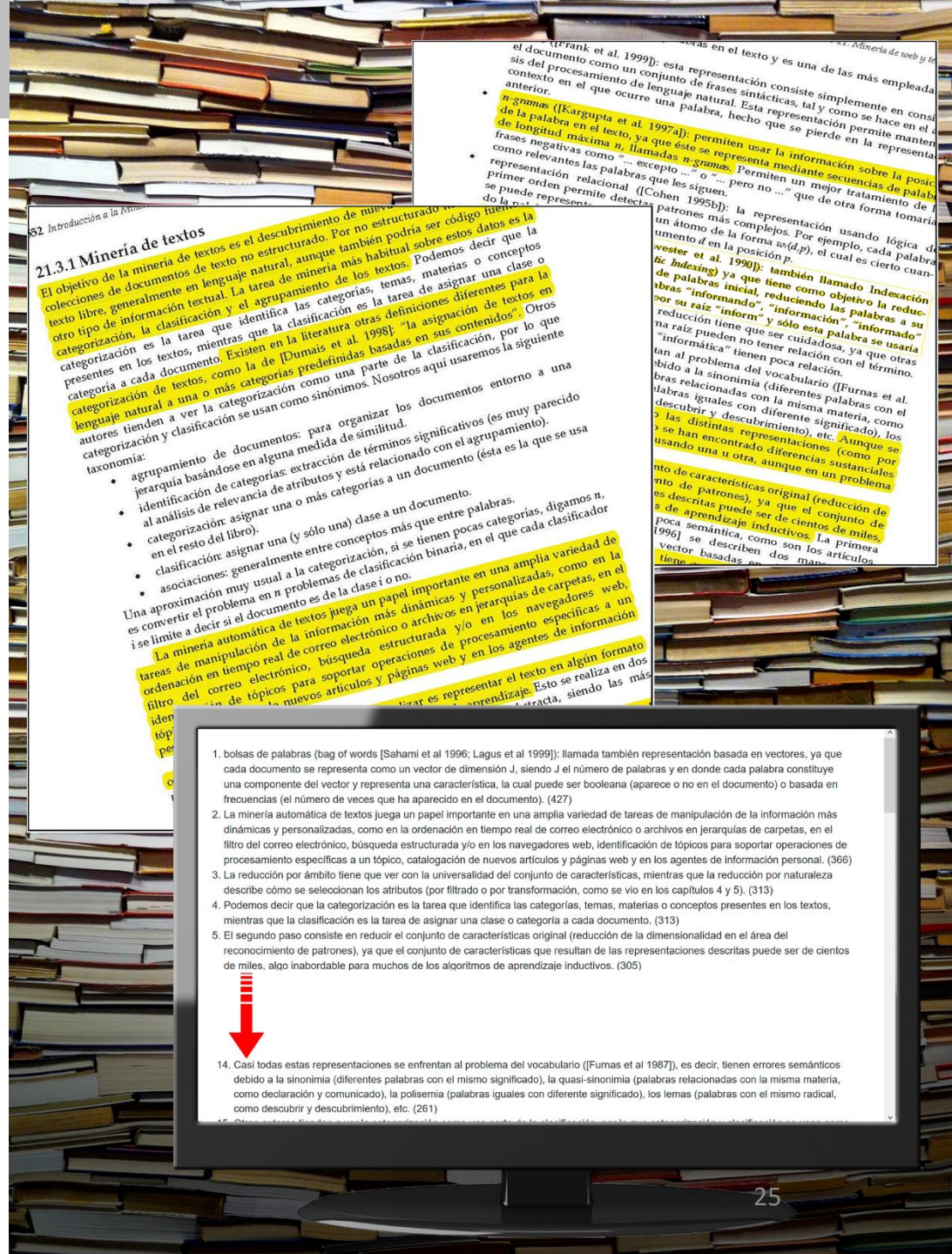


Resumen extractivo

21.3.1 Minería de textos

El objetivo de la minería de textos es el descubrimiento de nueva información a partir de colecciones de documentos de texto no estructurado. Por no estructurado nos referimos a texto libre, generalmente en lenguaje natural, aunque también podría ser código fuent tipo de información text tarea de minería más habitual sobre estos datos es la categorización, la clasificación y el agrupamiento de los textos. Podemos decir que la categorización es la tarea que identifica las categorías, temas, materias o conceptos presentes en los textos, mientras que la clasificación es la tarea de asignar una clase o categoría a cada documento.

POS_B

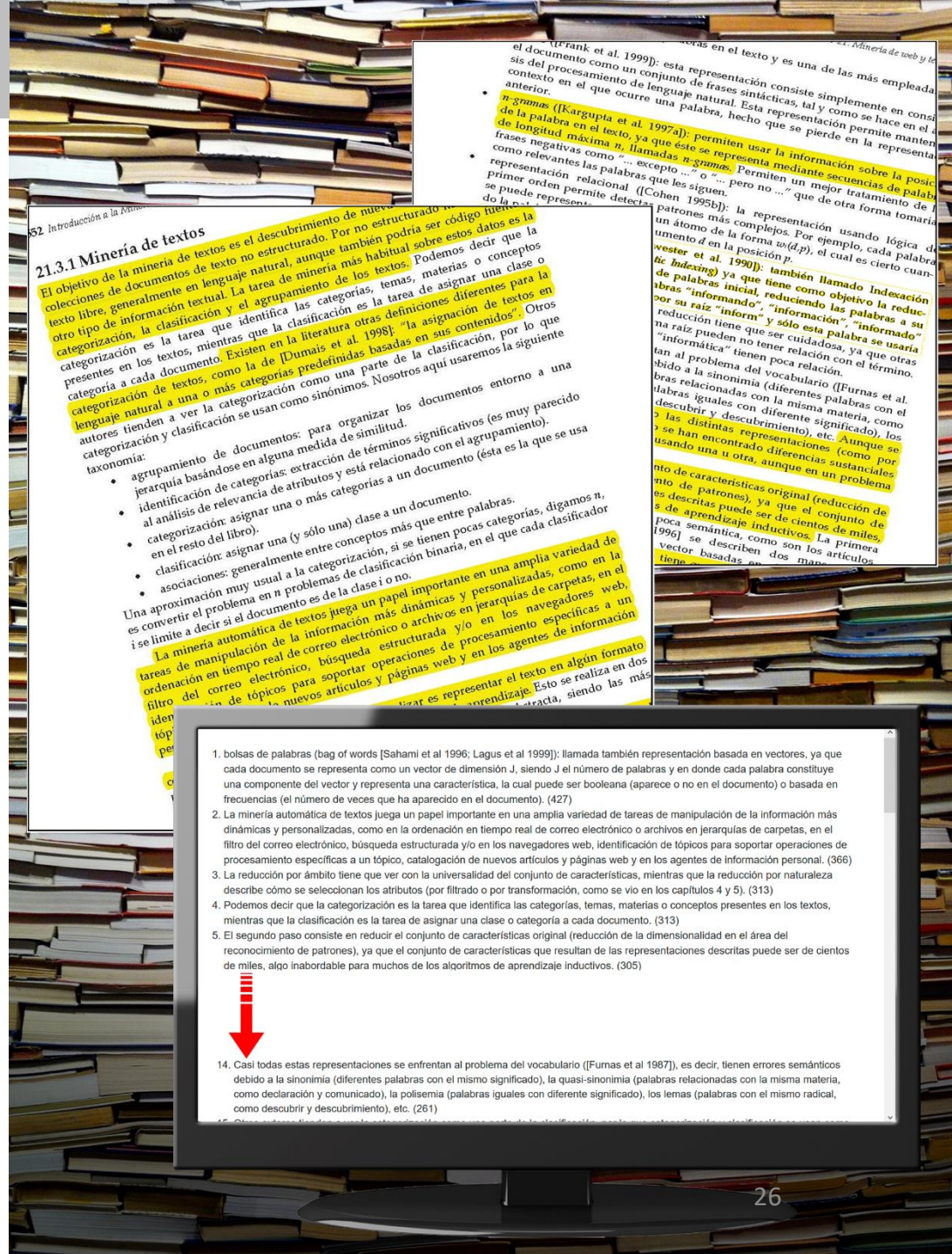


Resumen extractivo

TITLE

21.3.1 Minería de textos

El objetivo de la minería de textos es el descubrimiento de nueva información a partir de colecciones de documentos de texto no estructurado. Por no estructurado nos referimos a texto libre, generalmente en lenguaje natural, aunque también podría ser código fuente u otro tipo de información textual. La tarea de minería más habitual sobre estos datos es la categorización, la clasificación y el agrupamiento de los textos. Podemos decir que la categorización es la tarea que identifica las categorías, temas, materias o conceptos presentes en los textos, mientras que la clasificación es la tarea de asignar una clase o categoría a cada documento.

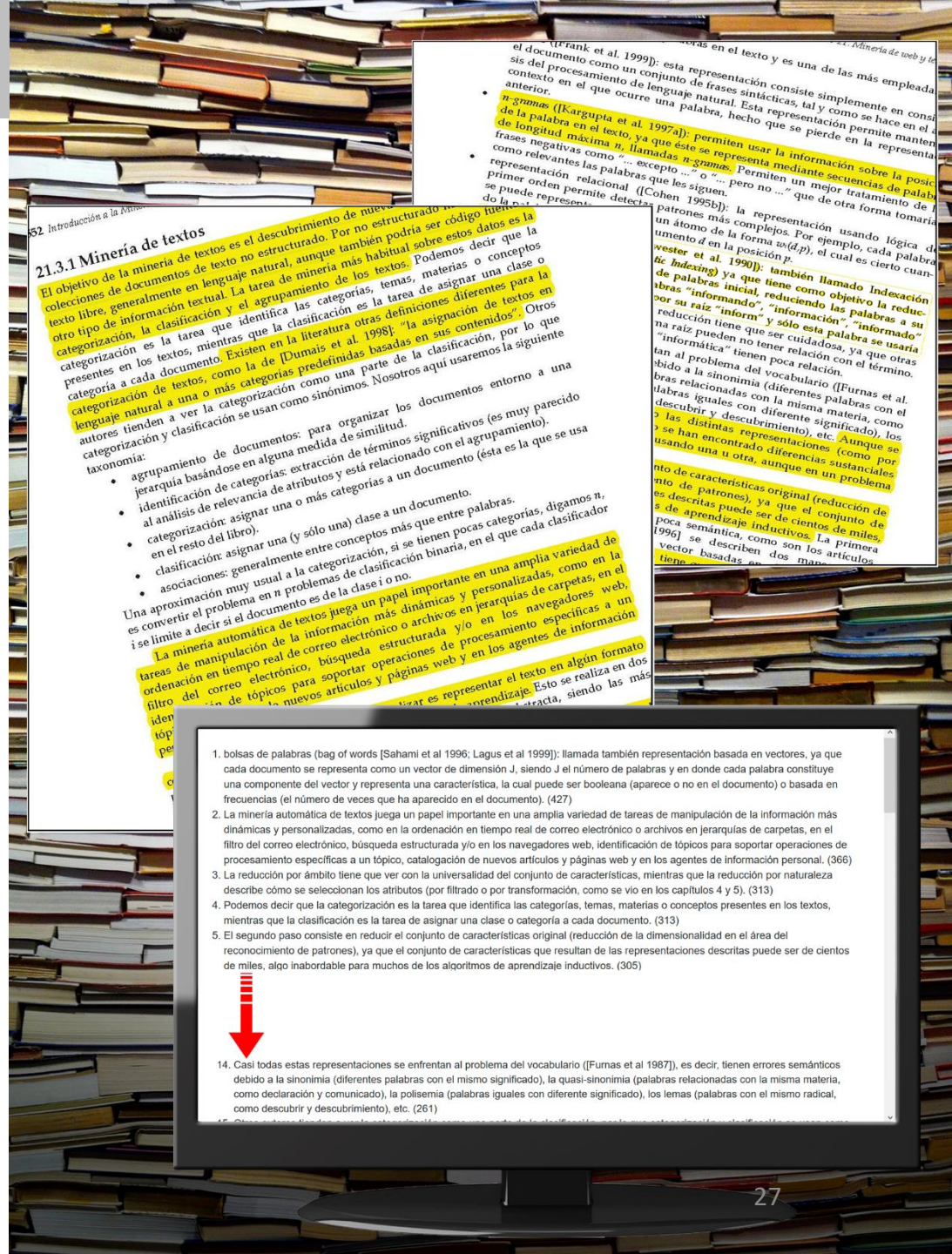


Resumen extractivo

21.3.1 Minería de textos

El objetivo de la minería de textos es el descubrimiento de nueva información a partir de colecciones de documentos de texto no estructurado. Por no estructurado nos referimos a texto libre, generalmente en lenguaje natural, aunque también podría ser código fuente u otro tipo de información textual. La tarea de minería más habitual sobre estos datos es la categorización, la clasificación y el agrupamiento de los textos. Podemos decir que la categorización es la tarea que identifica las categorías, temas, materias o conceptos presentes en los textos, mientras que la clasificación es la tarea de asignar una clase o categoría a cada documento.

LENGTH

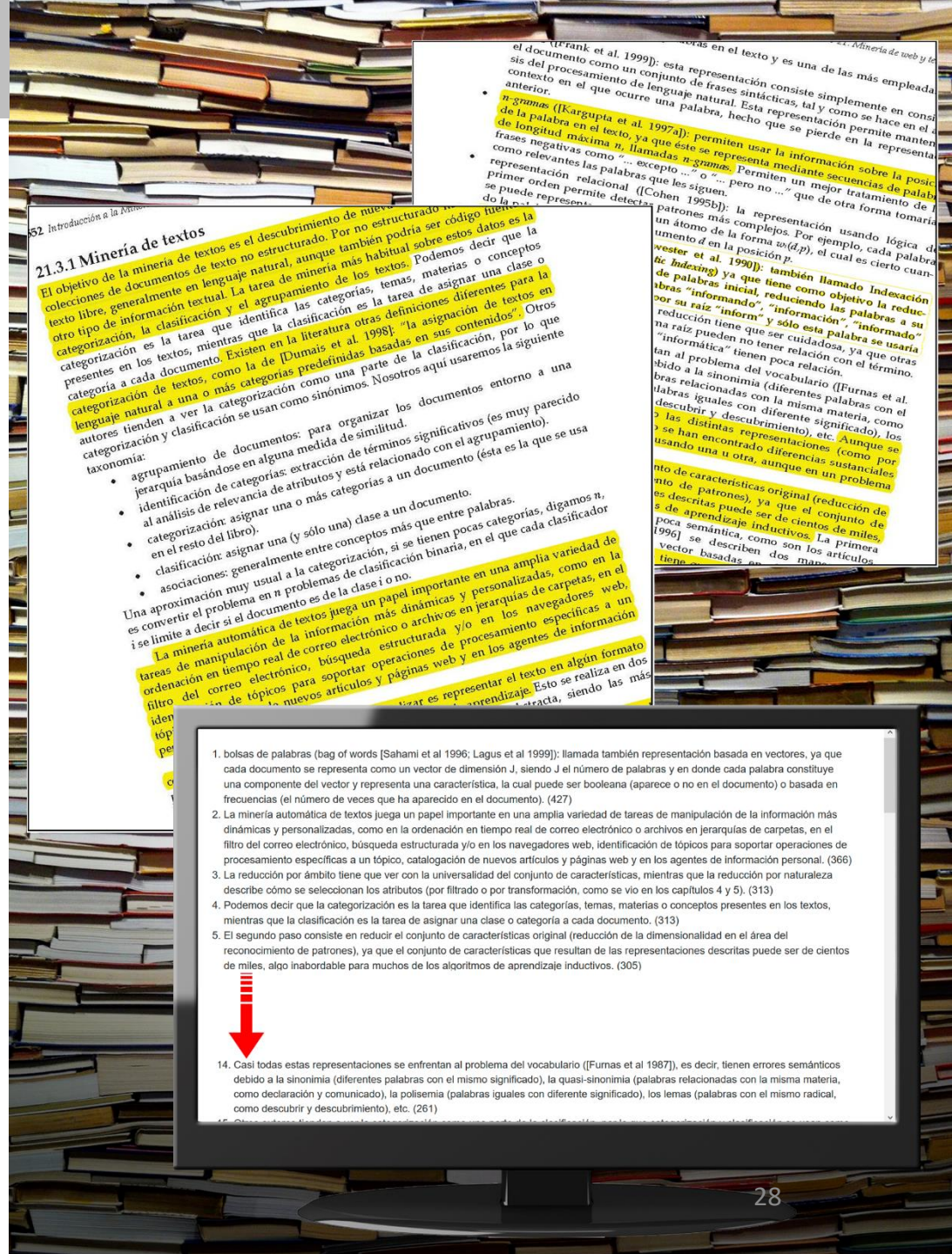


Resumen extractivo

21.3.1 Minería de textos

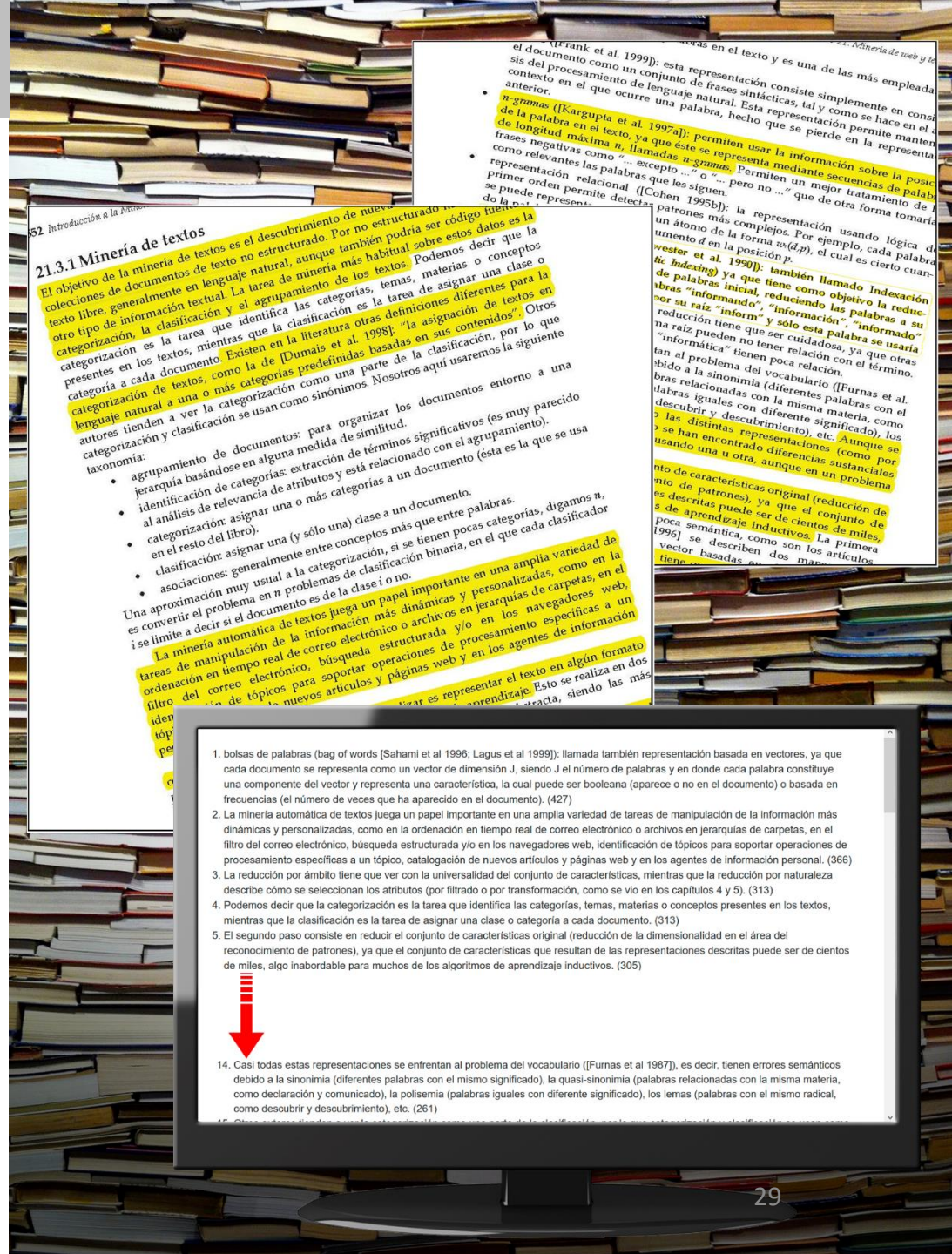
El objetivo de la minería de textos es el descubrimiento de nueva información a partir de colecciones de documentos de texto no estructurado. Por no estructurado nos referimos a texto libre, generalmente en lenguaje natural, aunque también podría ser código fuente u otro tipo de información [illegible] de minería más habitual sobre estos datos es la categorización, la clasificación y el agrupamiento de los textos. Podemos decir que la categorización es la tarea que identifica las categorías, temas, materias o conceptos presentes en los textos, mientras que la clasificación es la tarea de asignar una clase o categoría a cada documento.

MUCHAS OTRAS



Resumen extractivo

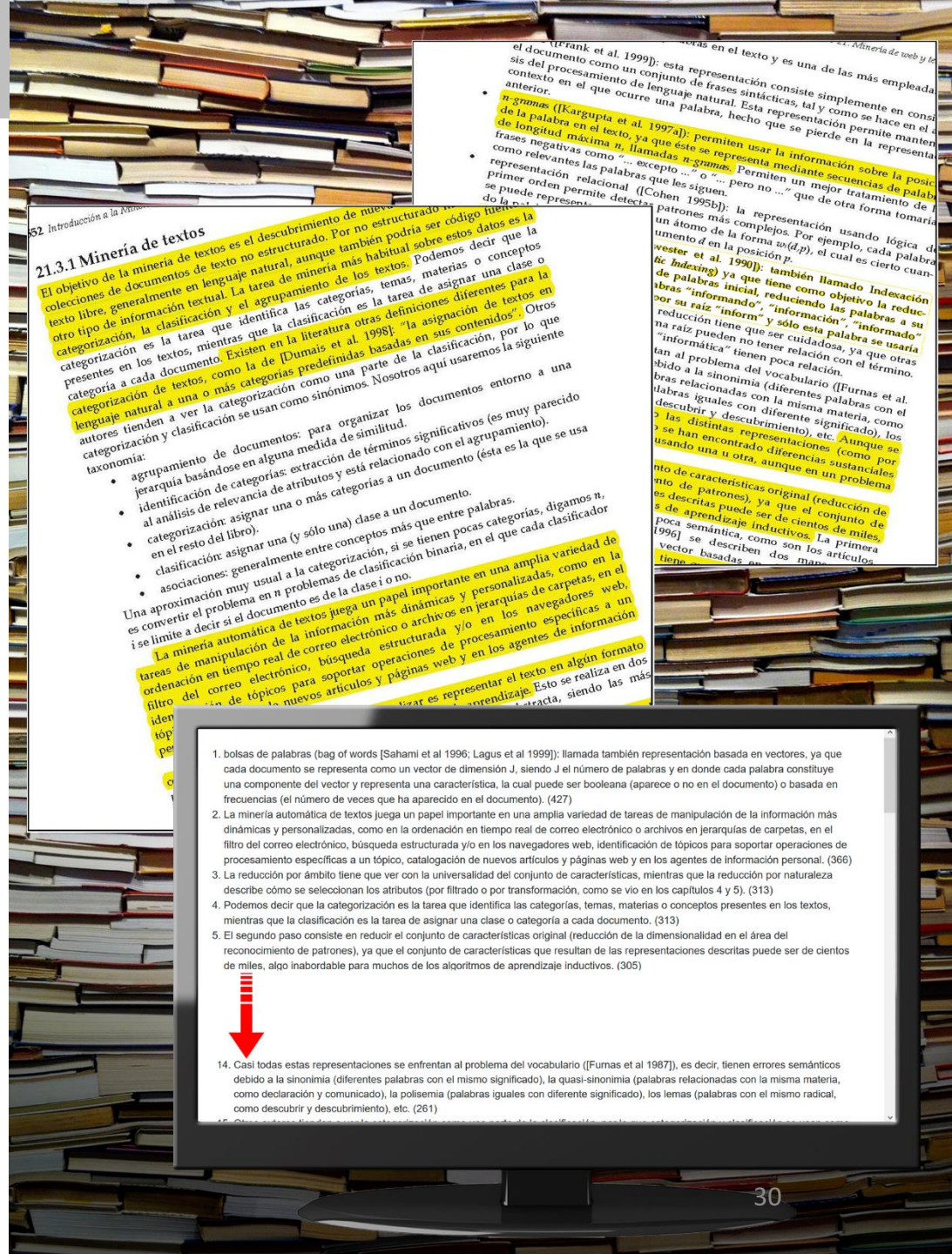
Tipo	Fórmula	Referencia
Posición	i i^{-1} $\max(i^{-1}, (p-i+1)^{-1})$	[Baxendale, 1958]
Longitud	$\lvert terms(s_i)\rvert$ $\sum_{t \in s_i} \lvert\{char : char \in t\}\rvert$	[Nobata et al., 2001]
Keywords	$\frac{\lvert keywords(c)\rvert^2}{\lvert c\rvert}$ $\sum_{k \in keywords(s_i)} \mathrm{tf}(k)$ $\frac{\lvert terms(s_i) \cap keywords(d_i)\rvert}{\lvert keywords(d_j)\rvert}$	[Luhn, 1958] [Edmundson, 1969] [Fatma et al., 2004]
Frecuencias	$\frac{\sum_{t \in s_i} \mathrm{tf}(t)}{\lvert terms(s_i)\rvert}$ $\sum_{t \in s_i} \mathrm{tf}(t) \cdot \mathrm{isf}(t)$	[Vanderwende et al., 2007] [Larocca Neto et al., 2000]
Títulos	$\frac{\lvert terms(s_i) \cap terms(titles)\rvert}{\min\left(\lvert terms(s_i)\rvert, \lvert terms(titles)\rvert\right)}$ $\frac{\lvert terms(s_i) \cap terms(titles)\rvert}{\lvert terms(s_i) \cup terms(titles)\rvert}$ $\frac{\vec{s_i} \times \vec{titles}(s_i)}{\lvert\vec{s_i}\rvert \times \lvert\vec{titles}(s_i)\rvert}$	[Edmundson, 1969]
Cobertura	$\frac{\lvert terms(s_i) \cap terms(d_j - s_i)\rvert}{\min\left(\lvert terms(s_i)\rvert, \lvert terms(d_j - s_i)\rvert\right)}$ $\frac{\lvert terms(s_i) \cap terms(d_j - s_i)\rvert}{\lvert terms(d_j)\rvert}$ $\frac{\vec{s_i} \times \vec{d_j - s_i}}{\lvert\vec{s_i}\rvert \times \lvert\vec{d_j - s_i}\rvert}$	[Litvak et al., 2010b]



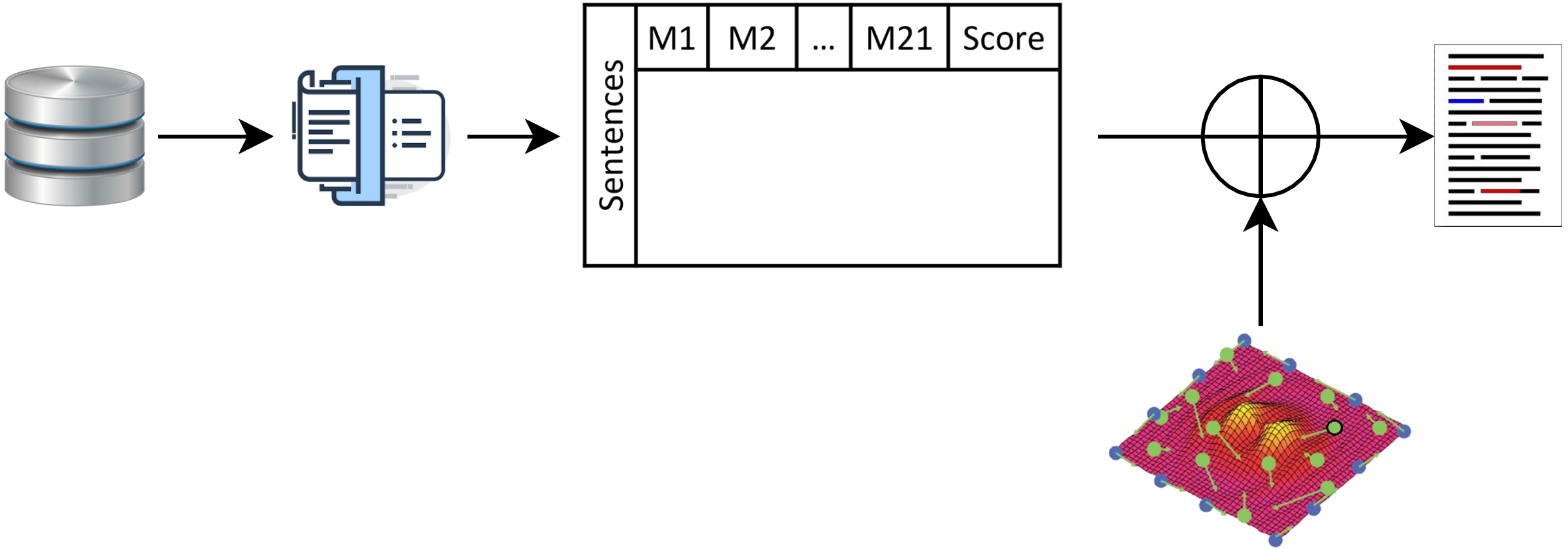
Resumen extractivo

El objetivo de la minería de textos es el descubrimiento de nueva información a partir de colecciones de documentos de texto no estructurado.

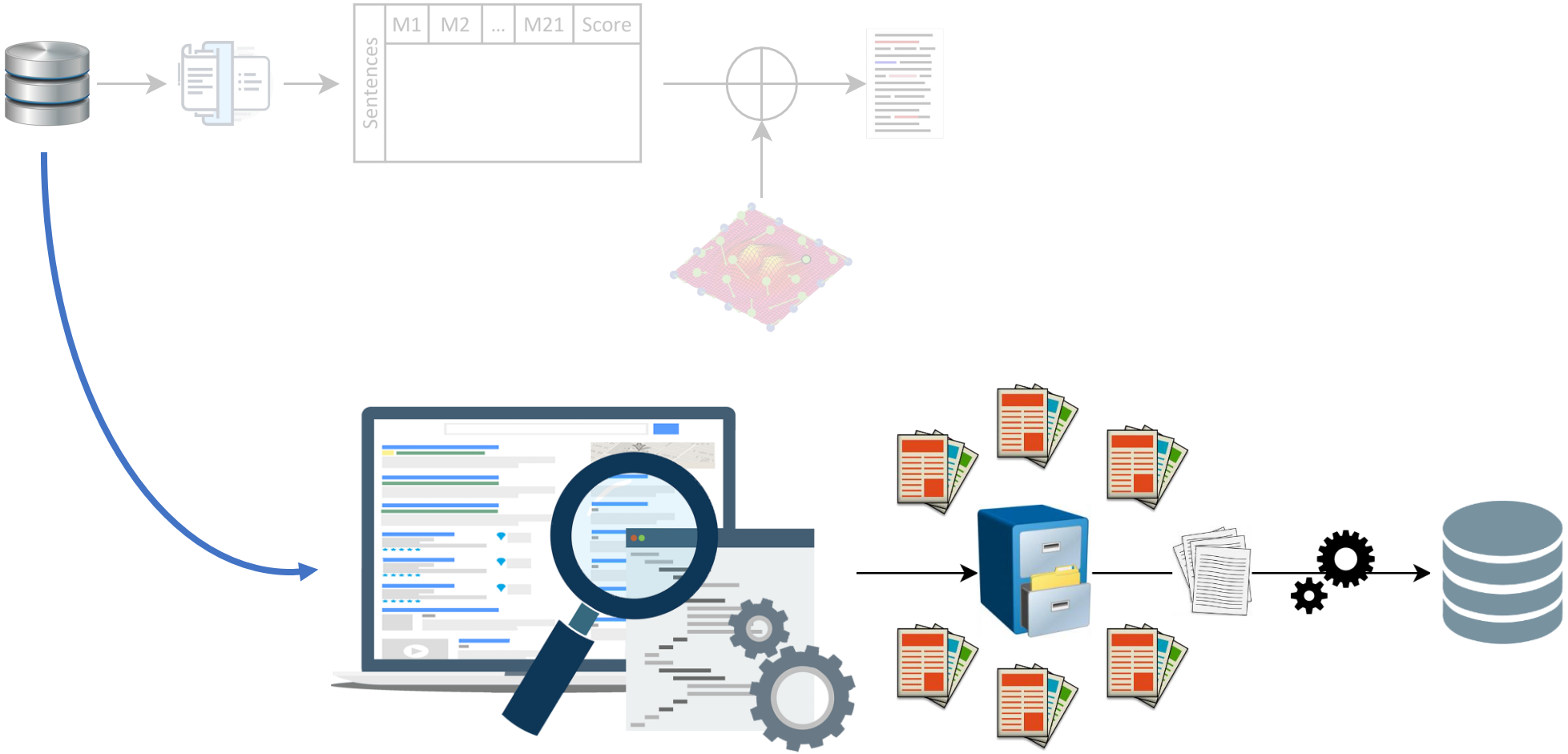
POS_L	LEN_CH	LUHN	KEY	TF	...	TITLE_O
1	0.37	0.13	0.09	0.56		0.67



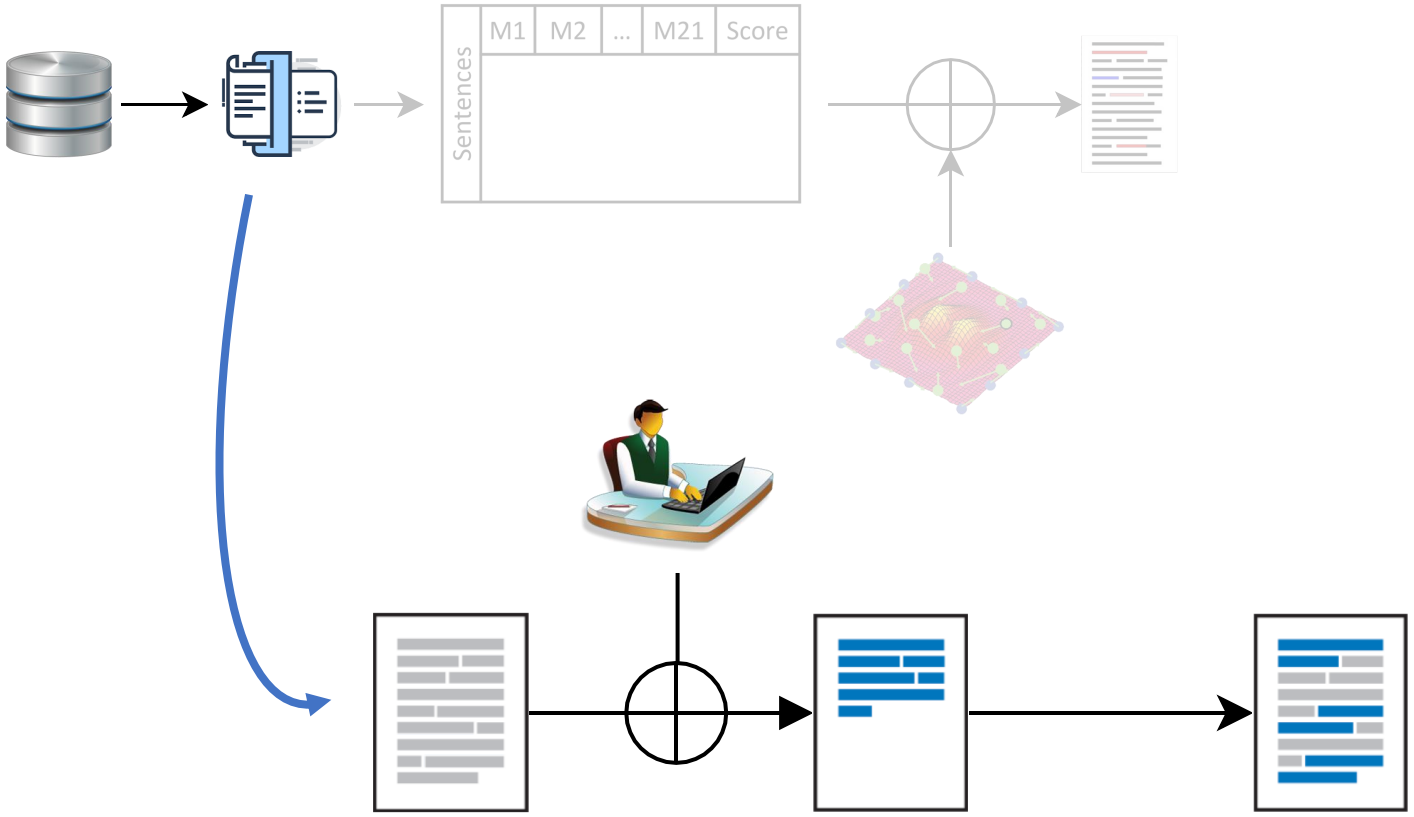
Resumen utilizando PSO



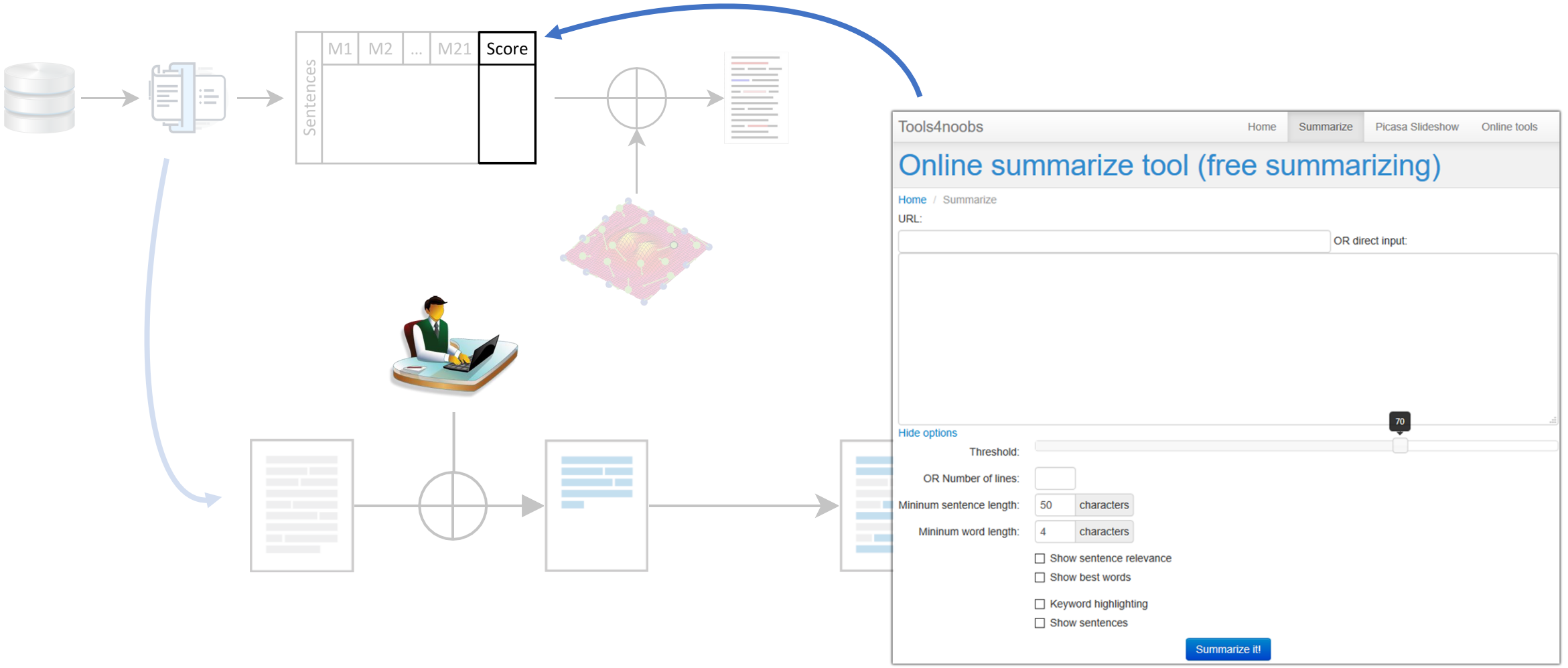
Resumen utilizando PSO



Resumen utilizando PSO

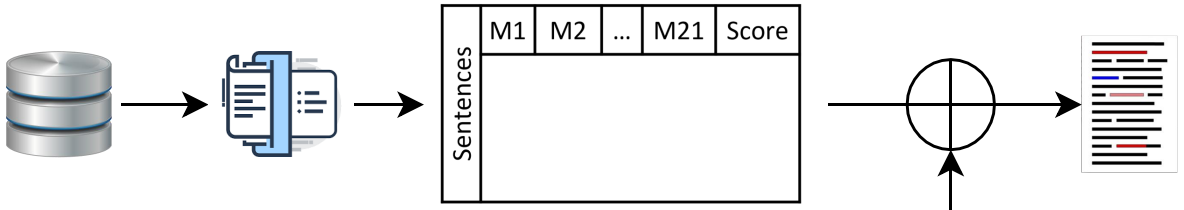


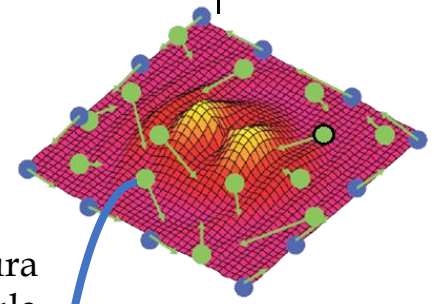
Resumen utilizando PSO



https://www.tools4noobs.com/summarize

Resumen utilizando PSO

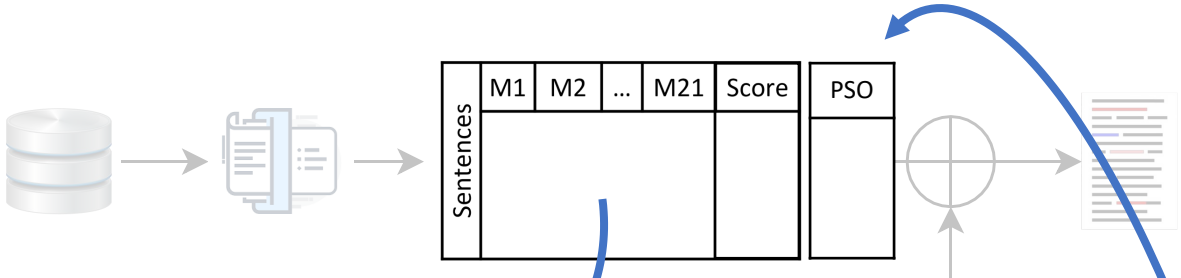




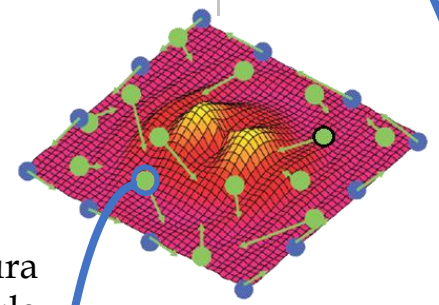
Estructura
de la partícula

	POS_L	POS_F	LEN_CH	LUHN	KEY	TF	...	TITLE_O
Parte binaria	0	1	0	1	0	1		1
Parte continua ($RealInd_i$)	1.26	0.26	-1.02	2.65	-0.69	1.06		1.32

Resumen utilizando PSO



$$S = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1n} \\ s_{21} & s_{22} & \cdots & s_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pn} \end{bmatrix}$$



Estructura de la partícula

$$fitness(p_i) = \sum_{d \in corpus} \frac{|summary(d)|}{|all\ summaries|} * MCC$$

$$score(RealInd_i, S_k) = \sum_{j=1}^{n} (realInd_{ij} * s_{kj})$$

	POS_L	POS_F	LEN_CH	LUHN	KEY	TF	...	TITLE_O
Parte binaria	0	1	0	1	0	1		1
Parte continua ($RealInd_i$)	1.26	0.26	-1.02	2.65	-0.69	1.06		1.32

Resumen utilizando PSO - Pruebas

Corpus de **3322 artículos** publicados en PLOS Medicine entre 2004 y 2018

- **Entrenamiento** con artículos de un mes y **testeo** sobre los del siguiente
- **30 ejecuciones** independientes con **100 iteraciones** como máximo
- **Resumen del 10%** del documento
- **PSO global de población fija** de 10 partículas inicializadas aleatoriamente
- **Comparación** del PSO que selecciona métricas con el que las utiliza a todas

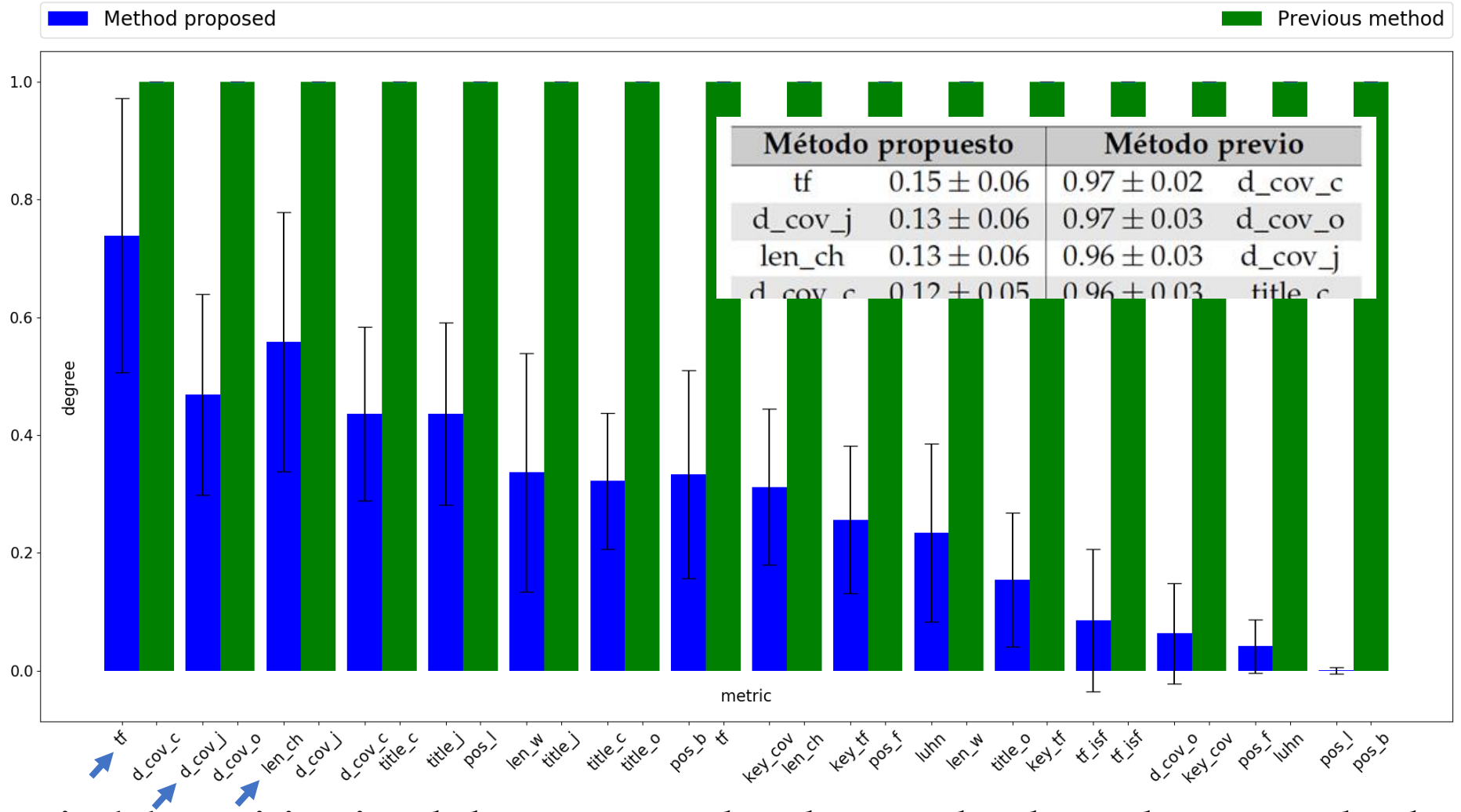
Resumen utilizando PSO - Resultados

Observemos las tres primeras

Método propuesto		Método previo	
tf	0.15 ± 0.06	0.97 ± 0.02	d_cov_c
d_cov_j	0.13 ± 0.06	0.97 ± 0.03	d_cov_o
len_ch	0.13 ± 0.06	0.96 ± 0.03	d_cov_j
d_cov_c	0.12 ± 0.05	0.96 ± 0.03	title_c
title_j	0.08 ± 0.04	0.95 ± 0.03	pos_l
len_w	0.08 ± 0.05	0.95 ± 0.03	title_j
title_c	0.06 ± 0.03	0.95 ± 0.03	title_o
pos_b	0.06 ± 0.04	0.95 ± 0.03	tf
key_cov	0.05 ± 0.03	0.95 ± 0.03	len_ch
key_tf	0.04 ± 0.03	0.95 ± 0.03	pos_f
luhn	0.04 ± 0.03	0.94 ± 0.04	len_w
title_o	0.03 ± 0.02	0.92 ± 0.04	key_tf
tf_isf	0.01 ± 0.02	0.92 ± 0.04	tf_isf
d_cov_o	0.01 ± 0.02	0.92 ± 0.04	key_cov
pos_f	0.01 ± 0.01	0.90 ± 0.05	luhn
pos_l	0.00 ± 0.00	0.89 ± 0.05	pos_b

Coeficientes promedio obtenidos con cada métrica. Estos valores corresponden a la media y desviación correspondiente.

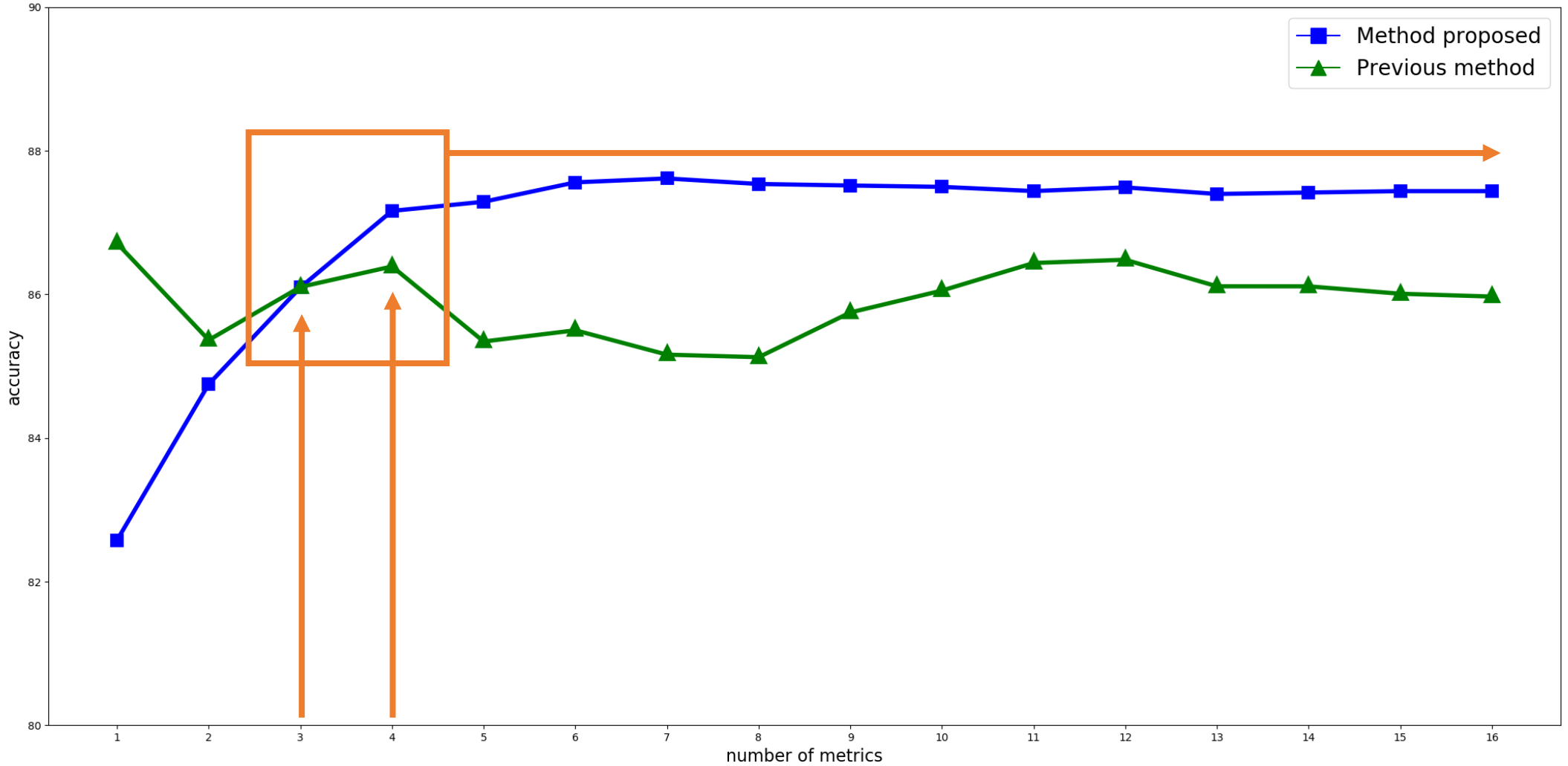
Resumen utilizando PSO - Resultados



Método propuesto		Método previo	
tf	0.15 ± 0.06	0.97 ± 0.02	d_cov_c
d_cov_j	0.13 ± 0.06	0.97 ± 0.03	d_cov_o
len_ch	0.13 ± 0.06	0.96 ± 0.03	d_cov_j
d_cov_c	0.12 ± 0.05	0.96 ± 0.03	title_c

Nivel de participación de las métricas ordenadas en orden descendente por valor de coeficiente.

Resumen utilizando PSO - Resultados



Evolución de la precisión a medida que se agregan nuevas métricas para calcular el puntaje.

Resumen utilizando causalidad y temporalidad

La **causalidad** cumple un **rol importante** en la cognición humana y en cualquier toda de decisiones

Ofrece reglas para explicar y predecir procesos complejos en términos de **relaciones "causa-efecto"**

- Describir **fenómenos**
- Responder **preguntas**
- Acompañar la respuesta de una **explicación**

Proceso **directo** (A causa B) o **indirecto** (A causa C a través de B)

Resumen utilizando causalidad y temporalidad

La **causalidad** cumple un **rol importante** en la cognición humana y en cualquier toda de decisiones

Ofrece reglas para explicar y predecir procesos complejos en términos de **relaciones "causa-efecto"**

- Describir **fenómenos**
- Responder **preguntas**
- Acompañar la respuesta de una **explicación**

Proceso **directo** (A causa B) o **indirecto** (A causa C a través de B)

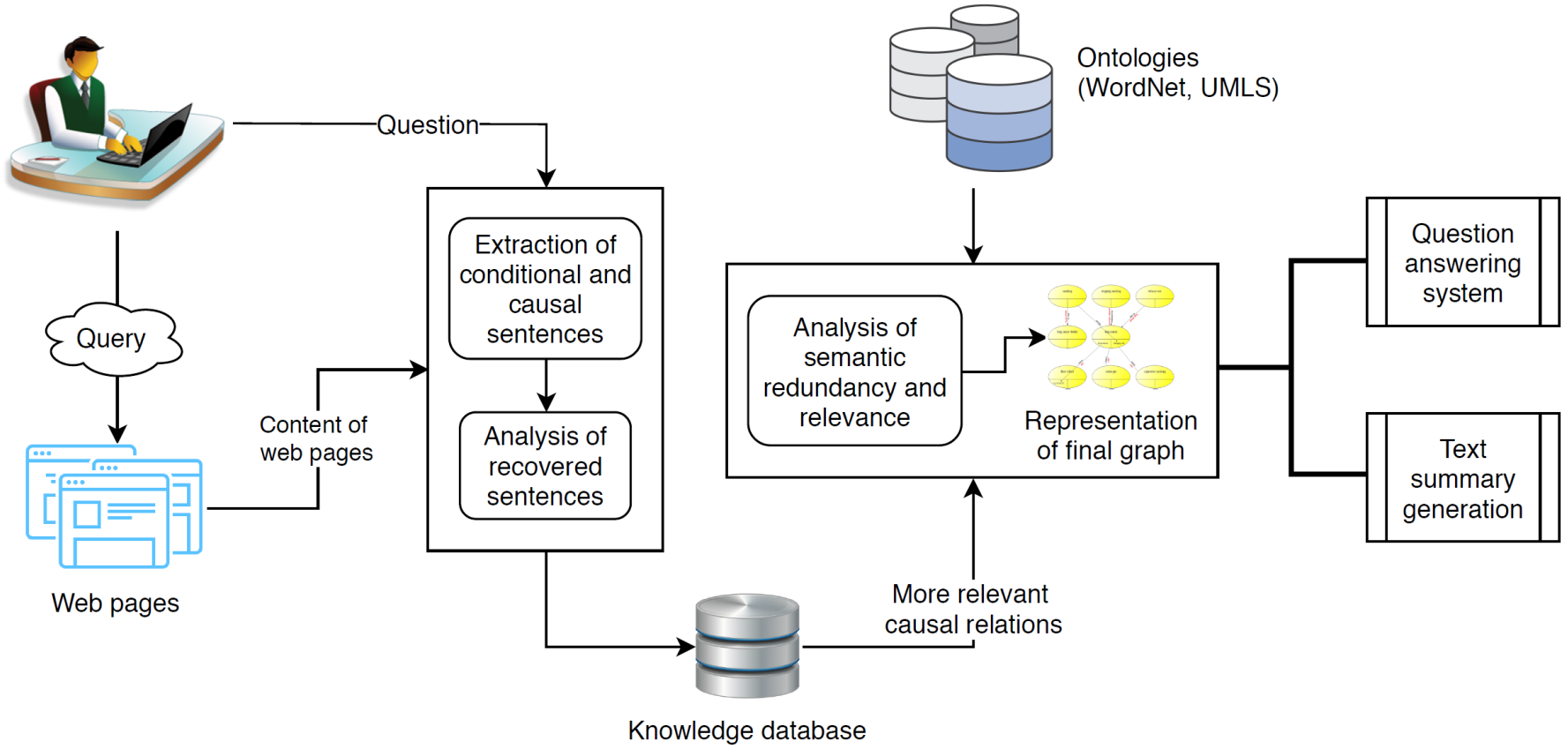
Resumen utilizando causalidad y temporalidad

1: if + present simple + future simple
2: if + present simple + may/might
3: if + present simple + must/should
4: if + past simple + would + infinitive
5: if + past simple + might/could
6: if + past continuous + would + infinitive
7: if + past perfect + would + infinitive
8: if + past perfect + would have + past participle
9: if + past perfect + might/could have + past participle
10: if + past perfect + perfect conditional continuous
11: if + past perfect continuous + perfect conditional
12: if + past perfect + would + be + gerund
13: for this reason, as a result
14: due to, owing to
15: provided that
16: have something to do, a lot to do
17: so that, in order that
18: although, even though
19: in case that, in order that
20: on condition that, supposing that

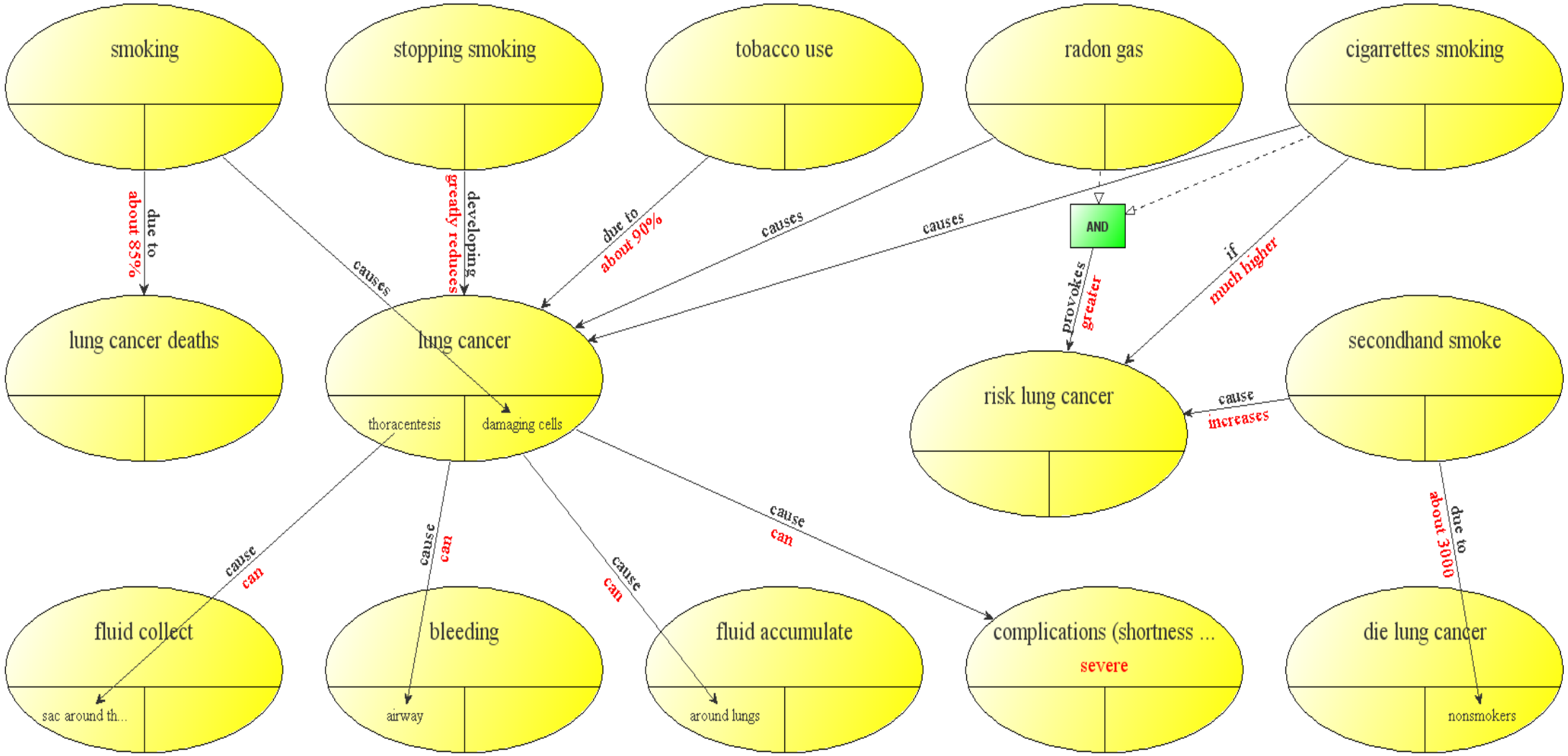
Estructuras causales y condicionales en inglés

Requiere un amplio conocimiento del lenguaje

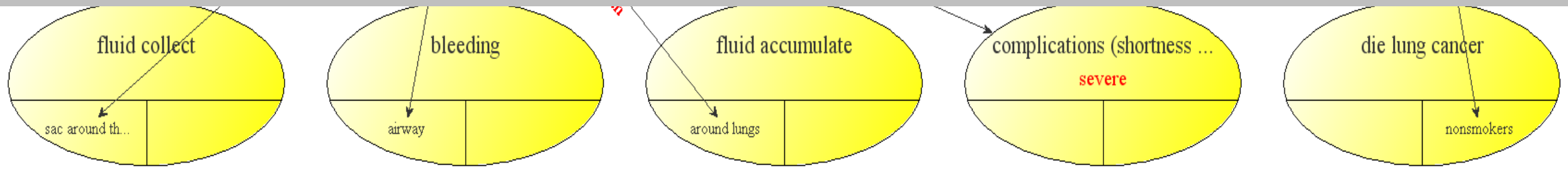
Resumen utilizando causalidad y temporalidad



Resumen utilizando causalidad y temporalidad



Resumen utilizando causalidad y temporalidad



"Cigarettes smoking causes die lung cancer occasionally and lung cancer normally. Tobacco use causes lung cancer constantly and die lung cancer infrequently. Lung cancer causes die lung cancer seldom and fluid collect sometimes. It is important to end knowing that lung cancer sometimes causes severe complication."

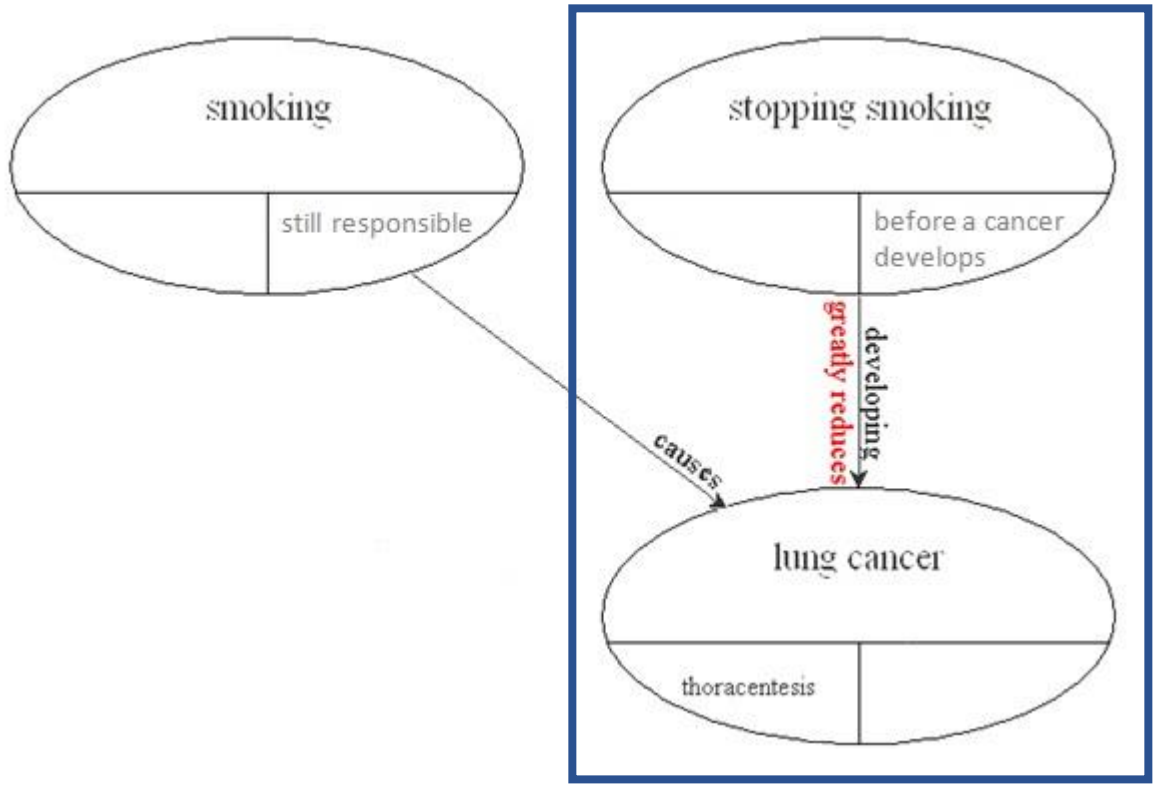
Las sentencias causales constituyen una parte importante de cualquier explicación médica

La toma de decisiones basada en el tiempo cobra un papel fundamental

www.elmundo.es/papel/historias/2018/05/11/5af43ad2468aeb486e8b

Resumen utilizando causalidad y temporalidad

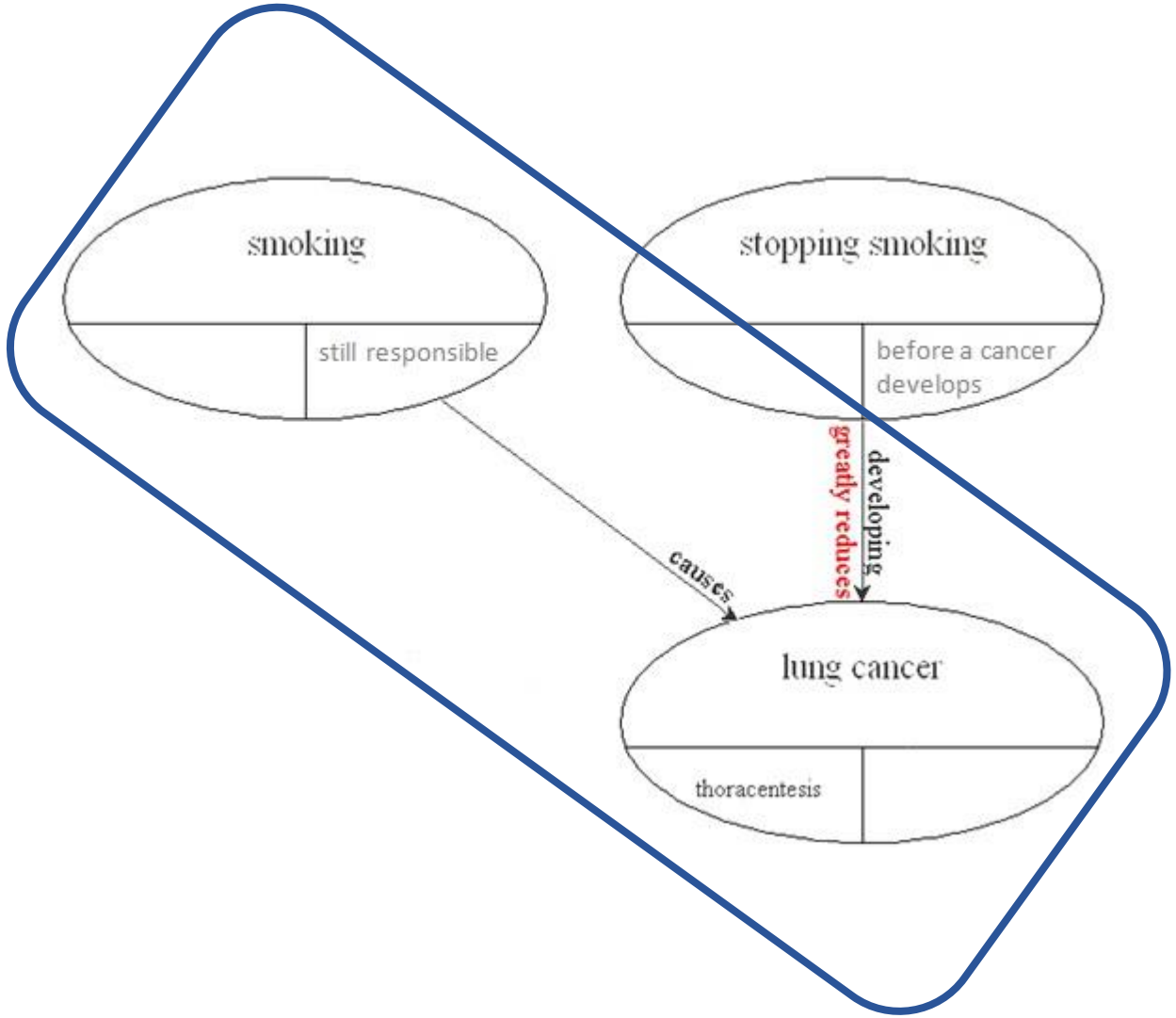
Restricción temporal *before* asociada a la causa *stopping smoking* (prerequisito temporal)



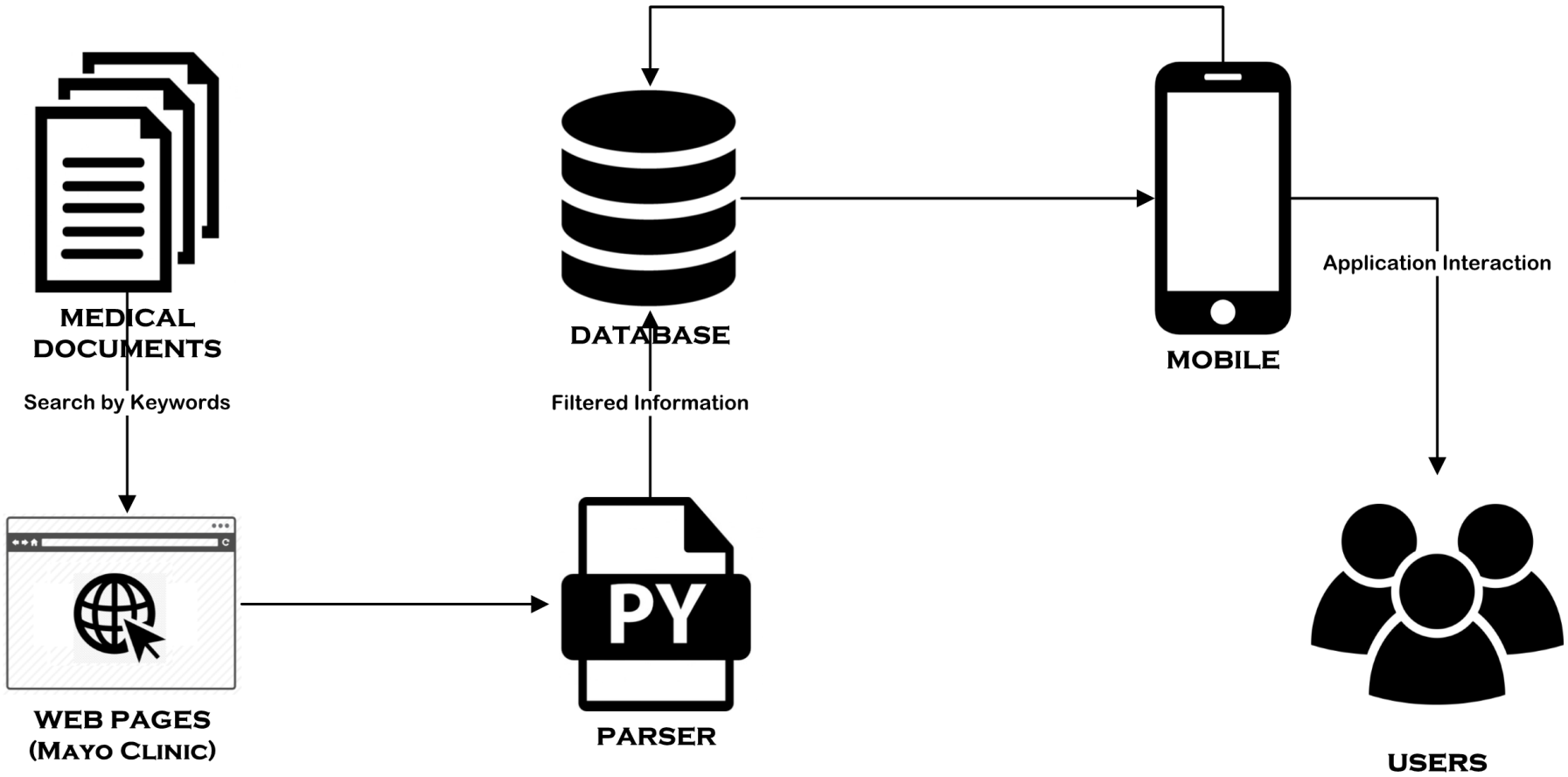
Resumen utilizando causalidad y temporalidad

Restricción temporal *before* asociada a la causa *stopping smoking* (prerequisito temporal)

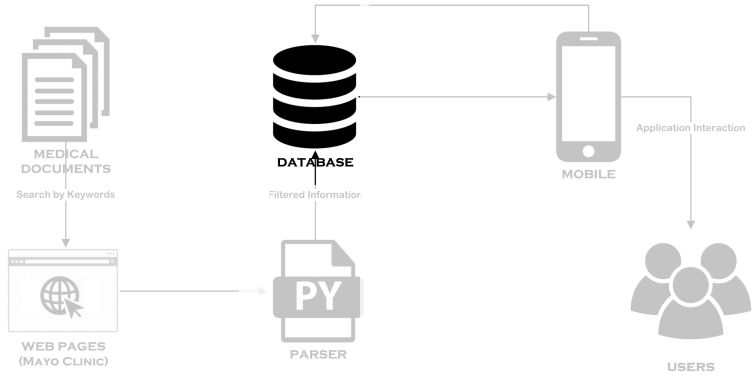
Componente temporal *still* no tiene restricción negativa hacia *lung cancer*

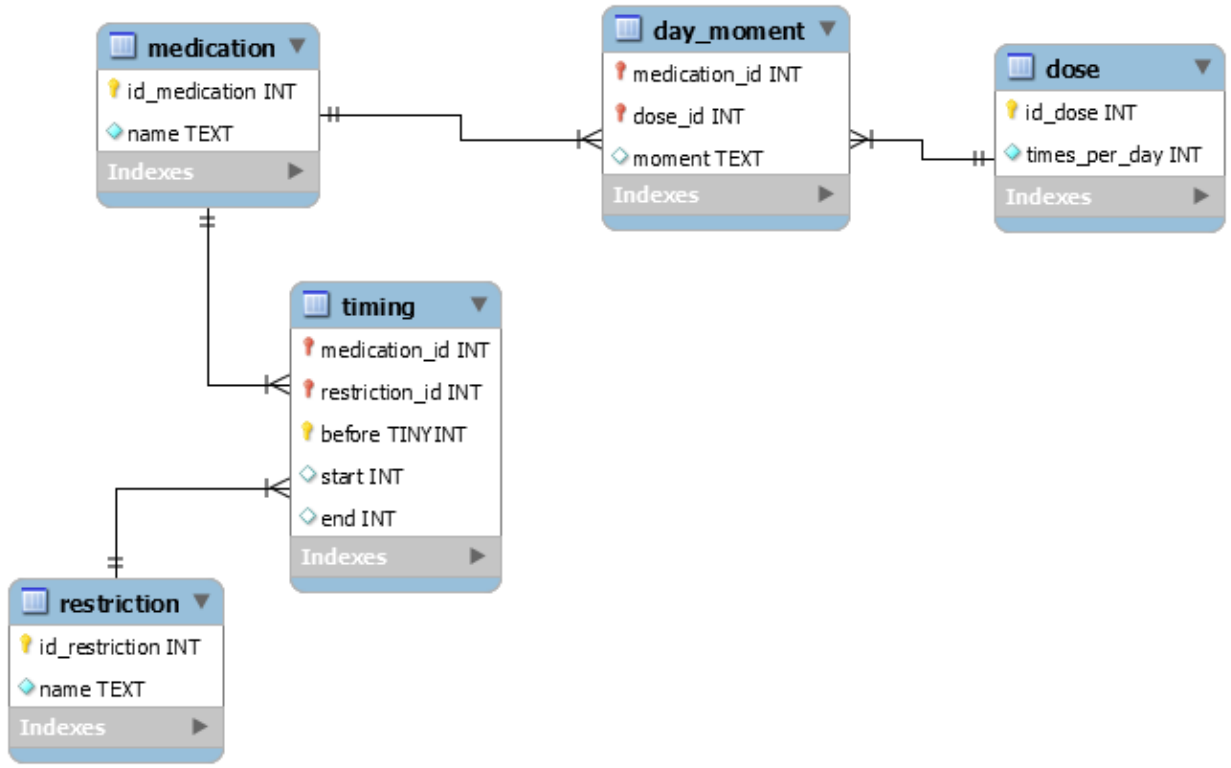


Proceso de control de administración de medicamentos

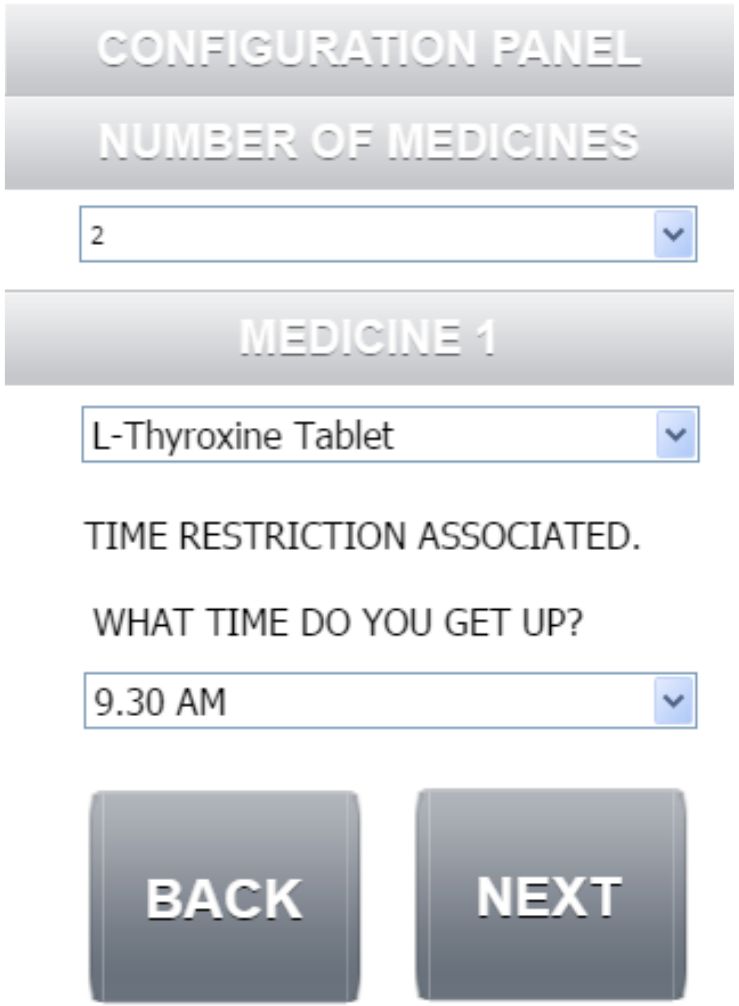


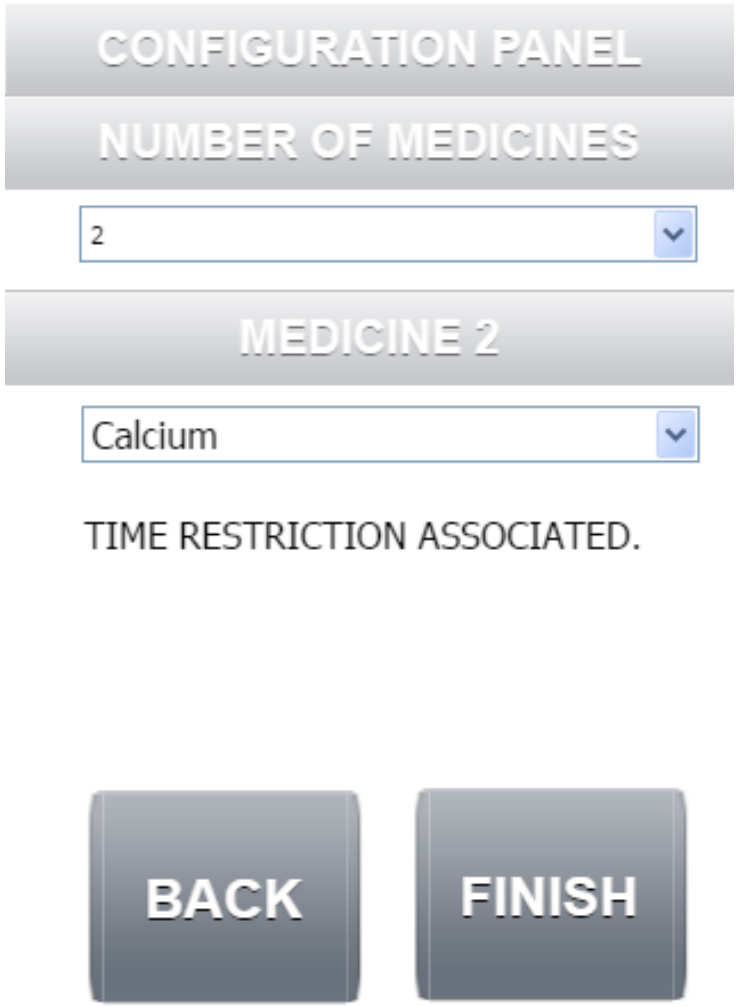
Proceso de control de administración de medicamentos





Aplicación para manejar restricciones temporales

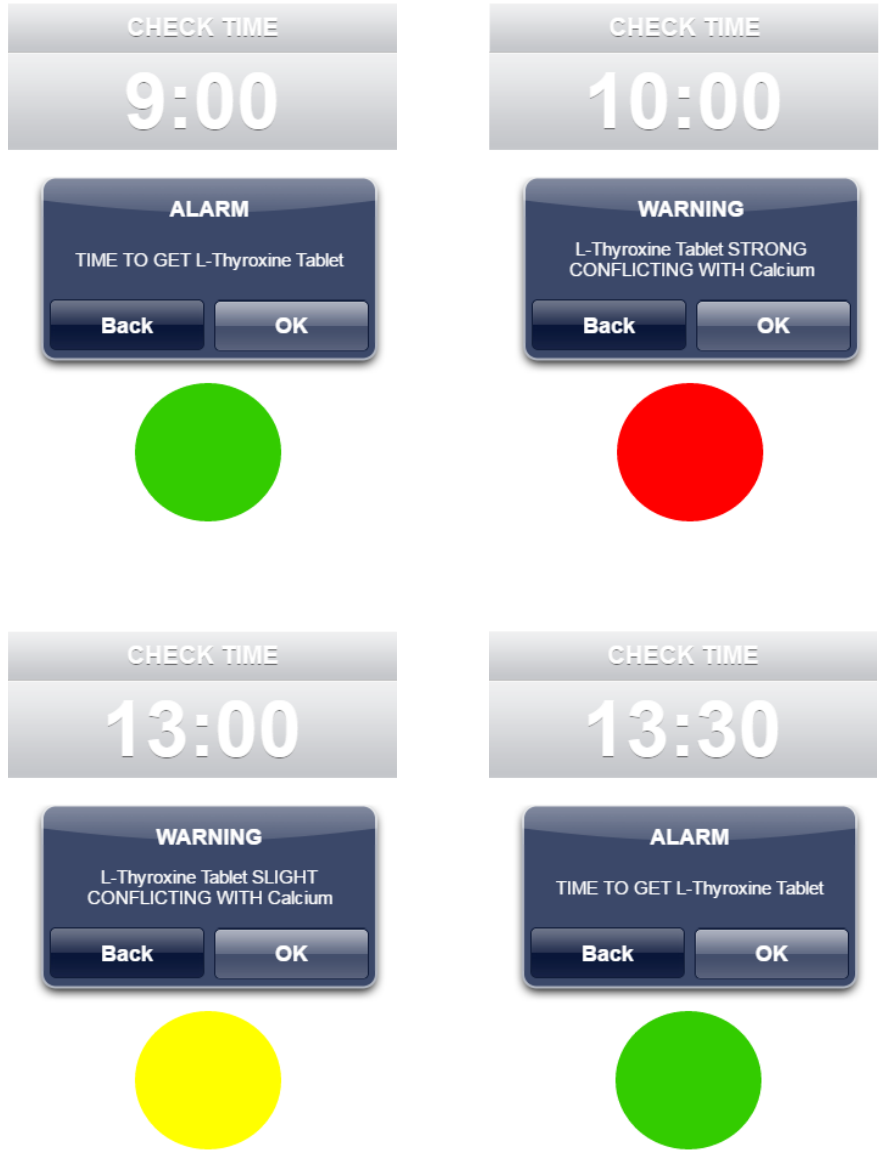




1. “Don’t take calcium supplements or antacids at the same time you take thyroid hormone replacement”

2. “Take any products containing calcium at least four hours before or after taking thyroid hormone replacement”

Pantallas de restricciones de tiempo



Conclusiones finales

- En este trabajo se han propuesto dos estrategias basadas en Inteligencia Artificial y en particular en técnicas de **Soft Computing** para la **generación automática de resúmenes de texto**

- Estas aproximaciones se han **evaluado** con **procedimientos habituales** en el área de estudio

- Han dado lugar a una **aplicación para dispositivos móviles**

Conclusiones finales

Propuestas:

- Método para **identificar el criterio del usuario** al seleccionar las partes principales de un documento por medio de una **variante original de PSO**
  - La técnica propuesta no se limita a resúmenes, habiendo sido utilizada en la obtención de reglas de clasificación

- Estrategia de extracción de **relaciones causales** en los textos a partir de las cuales construir un grafo (resumen abstractivo) con **anotaciones temporales** que afectan su interpretación

Conclusiones finales

Propuestas:

- Desarrollo de una aplicación que **combina la causalidad y temporalidad** para ayudar a controlar la administración de dosis de medicamentos

Conclusiones finales

Evaluación:

- La propuesta de **PSO** ha sido evaluada sobre una **amplia colección de artículos** científicos y los resultados reflejan que con **pocas métricas** es posible caracterizar el criterio del usuario para resumir

- Comparación de la calidad de los resúmenes generados de forma extractiva a partir de métricas y aquellos creados a partir de las **causales**. El resultado ha demostrado que la calidad está muy vinculada al **tipo de narrativa** del documento

Conclusiones finales

Evaluación:

- Los resultados de ambos métodos muestran que los resúmenes basado en **PSO** son más adecuados para **compactar el volumen de información** textual y los resúmenes abstractivos basados en **causales** son mejores para generar resúmenes **conceptualmente (semánticamente) más ricos**

Conclusiones finales

Cumplimiento de objetivos:

1. Explorar los aspectos claves de la obtención de resúmenes automáticos haciendo énfasis en los involucrados específicamente con el desarrollo de las soluciones propuestas
2. Analizar las tareas que intervienen en el preprocesamiento de texto e identificar cuáles permitirán representar adecuadamente los documentos en cada una de las soluciones.
3. Obtener el corpus de documentos a utilizar en cada caso y representarlo.

- Se han conseguido en un **alto grado**

Conclusiones finales

Cumplimiento de objetivos:

4. Diseñar y desarrollar las dos soluciones propuestas capaces de identificar a partir de los documentos de texto lo considerado relevante y por ende digno de ser conservado en el resumen final a construir.

- Se ha conseguido en un **alto grado** pero hay muchas otras cosas que se podrían incluir

Conclusiones finales

Cumplimiento de objetivos:

5. Determinar la relevancia del contenido causal de un documento y evaluar si es lo suficientemente importante como para formar un resumen extractivo.
6. Realizar experimentos y evaluar los resultados obtenidos de la aplicación de las soluciones desarrolladas.

- Se ha conseguido en un **alto grado** pero habría que hacer más pruebas de evaluación desde otros diferentes puntos de vista

Trabajos futuros

- **Ampliar el conjunto de métricas** utilizado para caracterizar los documentos de entrada y así enriquecer su representación
- Incorporar conceptos de **Lógica Borrosa o Difusa** que permitan flexibilizar el criterio del usuario y no utilizar valores exactos
- **Continuar con el desarrollo de la aplicación** para utilizarla además en la gestión hospitalaria
- Incluir **nuevas estrategias para verbalizar** el grafo causal resultante

Publicaciones

[1] Augusto Villa Monte, Laura Lanzarini, Luis Rojas Flores, and José A. Olivas. *Document summarization using a scoring-based representation*. In *2016 XLII Latin American Computing Conference*, pages 1−7, 2016.

[2] Cristina Puente, Augusto Villa Monte, Laura Lanzarini, Alejandro Sobrino, and José A. Olivas. *Evaluation of causal sentences in automated summaries*. In *2017 IEEE International Conference on Fuzzy Systems*, pages 1−6, 2017.

[3] Laura Lanzarini, Augusto Villa Monte, Aurelio F. Bariviera, and Patricia Jimbo Santana. *Simplifying credit scoring rules using LVQ+PSO*. *Kybernetes*, 46(1): 8−16, 2017.

[4] Augusto Villa Monte, Laura Lanzarini, Aurelio F. Bariviera, and José A. Olivas. *Obtaining and evaluation of extractive summaries from stored text documents*. In *Proceedings of the Third Conference on Business Analytics in Finance and Industry*, pages 65−66, 2018.

[5] Cristina Puente, Alejandro Sobrino, José A. Olivas, and Augusto Villa Monte. *Designing a system to extract and interpret timed causal sentences in medical reports*. *Journal of Experimental & Theoretical Artificial Intelligence*, 31(1): 1−13, 2019.

[6] Cristina Puente, Alejandro Sobrino, Augusto Villa Monte, and José A. Olivas. *Alert system for timely medication administration*. In *Proceedings of the 2018 International Conference on Artificial Intelligence (ICAI'18), located at 2018 World Congress in Computer Science, Computer Engineering, & Applied Computing (CSCE'18)*, pages 387−392. CSREA Press, 2018.

[7] Augusto Villa Monte, Julieta Corvi, Laura Lanzarini, Cristina Puente, Alfredo Simon Cuevas, and José A. Olivas. *Text pre-processing tool to increase the exactness of experimental results in summarization solutions*. In *Proceedings of the XXIV Argentine Congress of Computer Science*, 2018.

Publicaciones

[8] Augusto Villa Monte, César Estrebou, and Laura Lanzarini. *E-mail processing using data mining techniques*. In *Computer Science & Technology Series: XVI Argentine Congress of Computer Science - Selected papers*, pages 109−120. Edulp, 2011.

[9] Laura Lanzarini, Augusto Villa Monte, and César Estrebou. *E-mail processing with fuzzy SOMs and association rules*. *Journal of Computer Science and Technology*, 11(01): 41−46, 2011.

[10] Augusto Villa Monte, Franco Ronchetti, Laura Lanzarini, and Marcela Jerez. *Obtención de reglas de clasificación usando SOM+PSO*. In *Proceedings of the XVIII Argentine Congress of Computer Science*, pages 210−219, 2012.

[11] Laura Lanzarini, Augusto Villa Monte, Germán Aquino, and Armando De Giusti. *Obtaining classification rules using lvqPSO*. In *Advances in Swarm and Computational Intelligence*, volumen 9140 of *Lecture Notes in Computer Science*, pages 183−193. Springer International Publishing, 2015.

[12] Laura Lanzarini, Augusto Villa Monte, and Franco Ronchetti. *SOM+PSO: A novel method to obtain classification rules*. *Journal of Computer Science & Technology*, 15, 2015.

[13] Laura Lanzarini, Augusto Villa Monte, Aurelio F. Bariviera, and Patricia Jimbo Santana. *Obtaining classification rules using LVQ+PSO: An application to credit risk*. In *Scientific Methods for the Treatment of Uncertainty in Social Sciences*, volume 377 of Advances in Intelligent Systems and Computing, pages 383−391. Springer International Publishing, 2015.

[14] Patricia Jimbo Santana, Augusto Villa Monte, Enzo Rucci, Laura Lanzarini, and Aurelio F. Bariviera. *Analysis of methods for generating classification rules applicable to credit risk*. *Journal of Computer Science & Technology*, 17, 2017.

Gracias por su atención.