I Simposio Internacional

Tendencias en la Organización de la Información y del Conocimiento

SAN JOSÉ, COSTA RICA | 30-31 Oct y 01 Nov 2019

Indización Automática: una aproximación desde la inteligencia artificial

Francisco Castillo Chen
Ramón Masis Rojas
Daniel Morales Beita

01 Introducción

02 Indización: Indización Manual vs Indización Automática

03 Proyectos de Indización Automática

04 Modelo de IA con Inteligencia Artificial

05 Procesamiento del Lenguaje Natural

06 Demostración

Indización

Es un proceso de identificación del contenido de un documento y su descripción a través de términos verbales. La estrategia de organización de la información se basa en *descriptores* (palabras claves cuyos conceptos representan al documento en el que están contenidos) (Urbizastegui y Restrepo, 2011).

Determinar la materia temática de los documentos y expresarla en términos índices (descriptores, encabezamientos de materia, códigos de clasificaciones, términos indices), para hacer posible la recuperación temática (Mai, 2005).

Características Indización

- El conocimiento y familiaridad que tienen el indizador sobre el tema determinará el grado de consistencia que tiene el indizador.
- Es un proceso subjetivo ya que depende del conocimiento del indizador.
- Actualmente demanda altos costos para las bibliotecas y Centros de Información, ya que el proceso conlleva tiempo y recurso humano para largas horas de trabajo.
- El indizador debe ser capaz de interpretar uno o dos idiomas para realizar un procesos de calidad de la indización, así como conocer la terminologías usadas en cada idioma.
- Podemos encontrarla de dos tipos Indización **Manual e Indización Automática.**

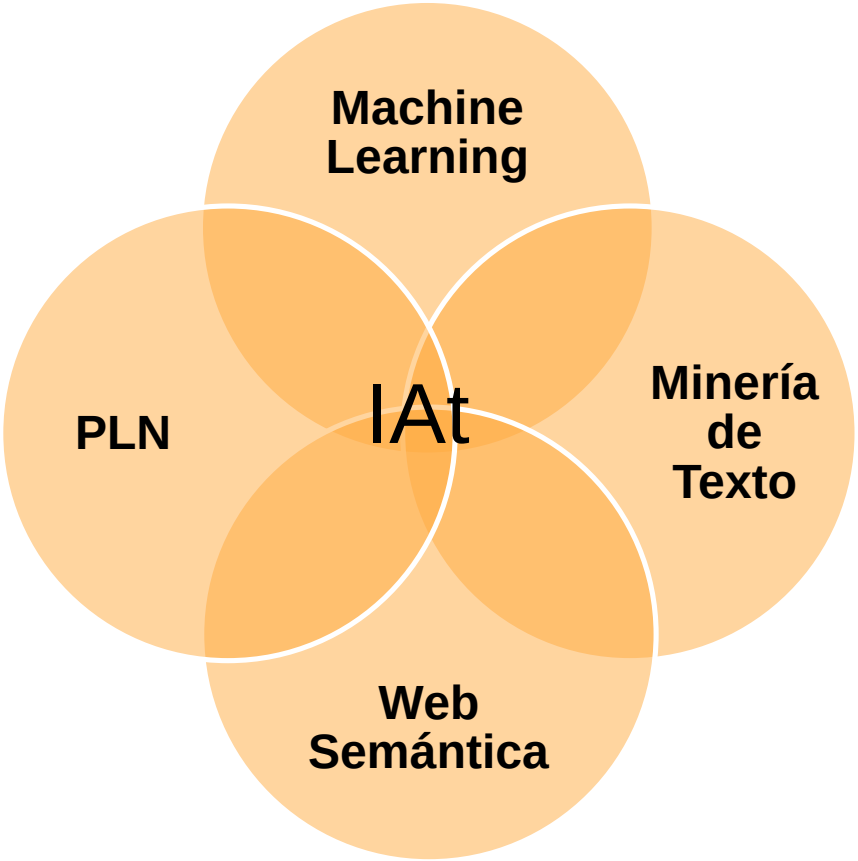
Indización Automática

Software informáticos que analizan, extraen y asignan a los documentos término de indización sin ninguna intervención humana (Gil Leiva, 2008). La indización automática es más rápida, económica, consistente y efectiva que la manual (Anderson, 1994)

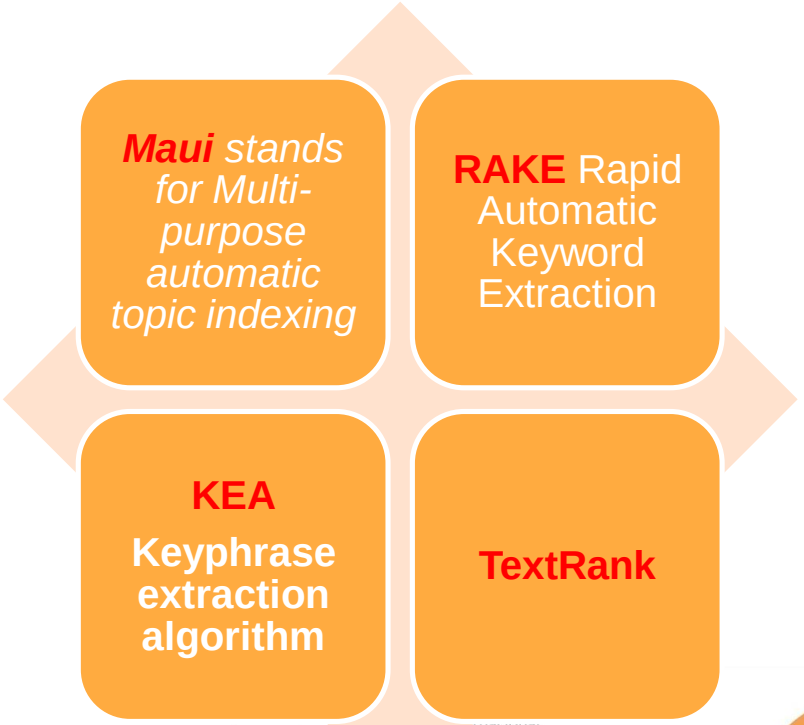
Un poco de Marco Teórico

- Luhn, H.P. (1958). The Automatic Creation of Literature Abstracts.
- Ribeiro, Lais A. (1974). Aplicação dos Métodos Estatísticos e da Teoria da informação e da Comunicação na Análise Linguística: estudo da linguagem jornalística
- Andreewski, A. y Vitoriano, Ruas. (1983). Indización automática basada en métodos lingüísticos y estadísticos y su aplicabilidad en lengua portuguesa
- Fugita, M. (1989). Avaliação da eficiência Mariangela Universidade da recuperação do S. L. Fugita Estadual sistema de indexação Paulista Precis
- Robrero, J. (1991). Indexação automática de textos: uma abordagem otimizada e simples.
- Pereira Flavia y Manfrim, B. (1991). Representação de conteúdo via indexação automática de textos integrais em língua portuguesa
- Gil, I (1996). Tendencia en la evolución de los sistemas de indización automática. Estudio Evolutivo.
- Urbizategui Alvarado, R.(1999). Aplicaciones de la Ley de Zipf a la indización Automática.
- Severina Mendonça, E. (2000). A linguística e a ciência da informação: estudos de uma interseção
- Jimenez Salazar, H.;Pinto, D.; y Rosso, P. (2005).Uso del punto de transición en la selección de términos índice para agrupamiento de textos cortos.
- Urbizategui Alvarado, R.(2011). La Ley de Zipf y el punto de transición de Goffman, en la indización automática.
- Precisão no processo de busca e recuperação da informação: uso da mineração de textos
- Gil, I. (2017) SISA—Automatic Indexing System for Scientific Articles: Experiments with Location Heuristics Rules versus TF-IDF Rules [Sistema de Indizacion Automática para articulos científicos]

Relación de la Indización Automática con otras disciplinas



Algunos algoritmos para el uso de la indización automática



Ejemplo de uso del algoritmo KEA en indización automática con el Tesauro Agrovoc de FAO

The growing global obesity problem: Some policy options to address it		Overview of techniques for reducing bird predation at aquaculture facilities		Feeding Asian cities: Food production and processing issues	
Indexers	Kea-4.0	Indexers	Kea-4.0	Indexers	Kea-4.0
developing countries	developing countries	aquaculture	aquaculture	food policies	food policies
food consumption	food consumption	damage	damage	food supply	food supply
overweight	overweight	fencing	fencing	towns	towns
taxes	taxes	noise	noise	urbanization	urbanization
prices	price fixing	scares	scares	urban agriculture	urban agriculture
price policies	controlled prices	bird control	birds	urban areas	urban areas
fiscal policies	policies	fishery production	fishing operations	food production	food consumption
feeding habits	body weight	predatory birds	predators	agricultural policies	new products
nutritional requirements	saturated fats	control methods	ropes	rural urban relations	
diet		fish culture		asia	
nutrition policies		noxious birds		state intervention	
food intake				agricultural sector	
				suburban agriculture	

	Indizador	KEA 4,0	Similtud de terminos
Articulo 1	**12**	**9**	**5**
Articulo 2	**11**	**9**	**5**
Articulo 3	**13**	**8**	**6**

Como puede observarse el algoritmo utilizado presenta muchas coincidencias con los términos utilizados por los indizadores profesionales

Proyectos de Indización Automática

En la actualidad a nivel mundial la principal Biblioteca que lidera este proceso y lo viene implementando en diversa medida es la National Library of Medicine- National Institute of Health (NLM https://ii.nlm.nih.gov/ de los Estados Unidos). **Indexing Iniciative II**

Objetivo: Investigar métodos basados en el lenguaje y el aprendizaje automático para la selección automática de encabezamientos de materia para su uso en entornos de indización semiautomatizados y totalmente automatizados en NLM.

Proyectos de Indización Automática

A nivel Iberoamericano España recientemente lanzo el proyecto (2013): **PLANTL (Plan de Impulso de las Tecnologías del Lenguaje).** Este proyecto pretende crear plataformas comunes de procesamiento de lenguaje natural para las Administraciones Públicas.

Paralela a esta iniciativa se desarrolla la **evaluación para la indización semántica de literatura medica en español** el cual tiene como objetivo fomentar el desarrollo del procesamiento del lenguaje natural y la **traducción automática en lengua española y lenguas co-oficiales**

1. Barcelona Supercomputing Center
2. Centro Nacional de Investigaciones Oncológicas
3. Biblioteca Nacional de Ciencias de la Salud
4. BIOASQ (http://bioasq.org/)

LILACS (Base de datos de Literatura Latinoamericana en Ciencias de la Salud),
IBECS (Índice Bibliográfico Español en Ciencias de la Salud).

Demostración del Corpus:
https://www.plantl.gob.es/tecnologias-lenguaje/actividades/demostradores/Paginas/corpus-viewer.aspx

PLANTL: https://www.plantl.gob.es/tecnologias-lenguaje/actividades/plataformas/Paginas/plataformas-pln.aspx

Inteligencia Artificial

La ciencia llamada inteligencia artificial (IA, 1956), **se ha dado por objetivo el estudio y el análisis del comportamiento humano**. De esta manera, las aplicaciones de la IA **se sitúan principalmente en la simulación de actividades intelectuales del hombre.** Es decir, imitar por medio de máquinas, normalmente electrónicas, tantas actividades mentales como sea posible, y quizás llegar a mejorar las capacidades humanas en estos aspectos.

Inteligencia Artificial-Tipologias

- Aquellos que sostienen que es posible realizar “dispositivos realmente pensantes”, punto de vista llamado la **IA fuerte**.

- Los que sostienen que es posible simular estados mentales de nuestro cerebro por medio de computadores, punto de vista llamado **IA débil**.

- y finalmente los “dualistas”, quienes dan por separada – muy en resumidas cuentas - la dimensión del cuerpo y del espíritu, y que de esta manera, existirán “juicios de verdad” a los cuales las computadores no tendrán nunca acceso.

Consultado en: **journals.openedition.org (2019)**

Aplicaciones de la Inteligencia Artificial

NACIONALES

Adoptar tecnología aumentaría productividad y llevaría el PIB al 7,8% en 10 años

Inteligencia artificial salvaría la economía costarricense

País debe realizar profundos cambios en materia empresarial, educativa y social, según estudio de Microsoft y DuckerFrontier

Inteligencia Artificial

Gráfico No.1
Cantidad de articulos publicados en *Scielo Citation Index* referente a Inteligencia Artificial, 2019

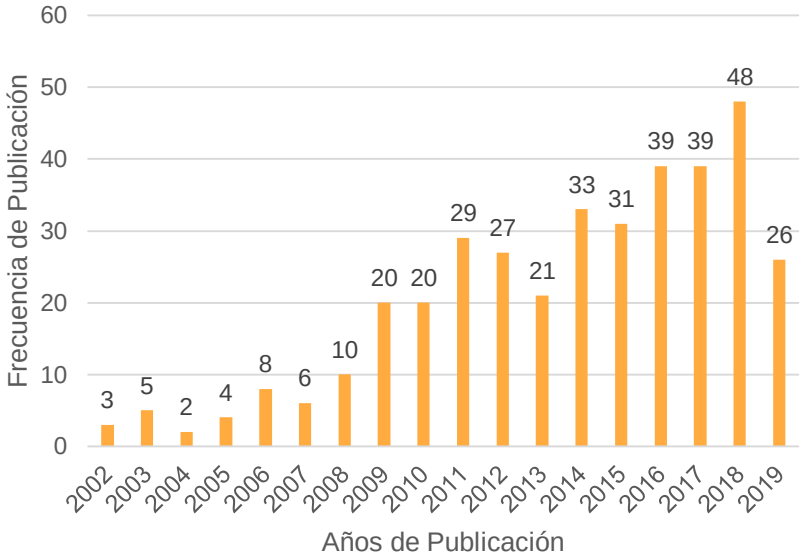
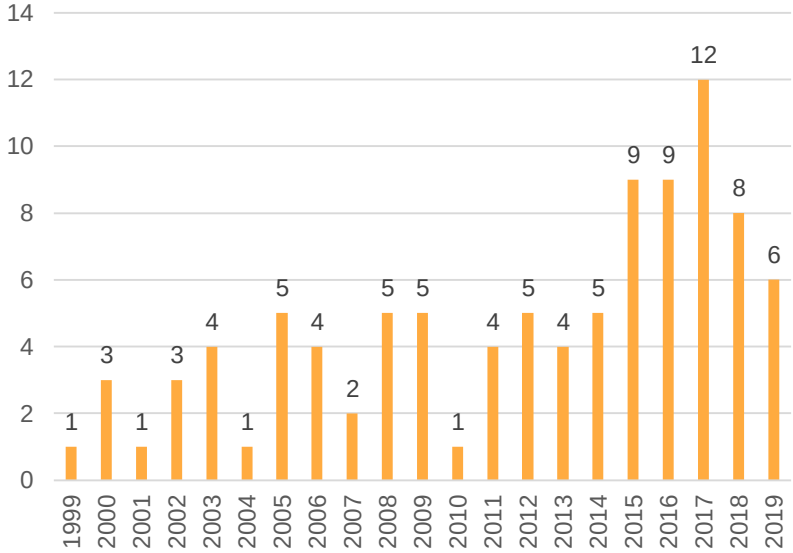


Gráfico No. 2
Cantidad de articulos publicados en Web of Science Colección Principal referente a Inteligencia Artificial y Bibliotecología/Bibliotecas, 2019



Proceso de Indización

Políticas de Indización

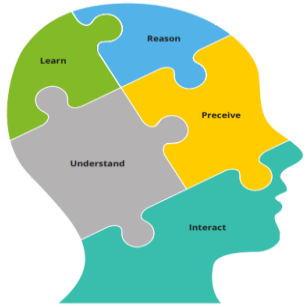
Objetivo de la indización

Complejidad del Objeto indizado

Conocimiento de la materia

Dominio de las herramientas de indización

INDIZADOR



DOCUMENTO

TITULO

RESUMEN / INTRODUCCIÓN

TABLA DE CONTENIDO

CONTENIDO

DESCRIPTORES

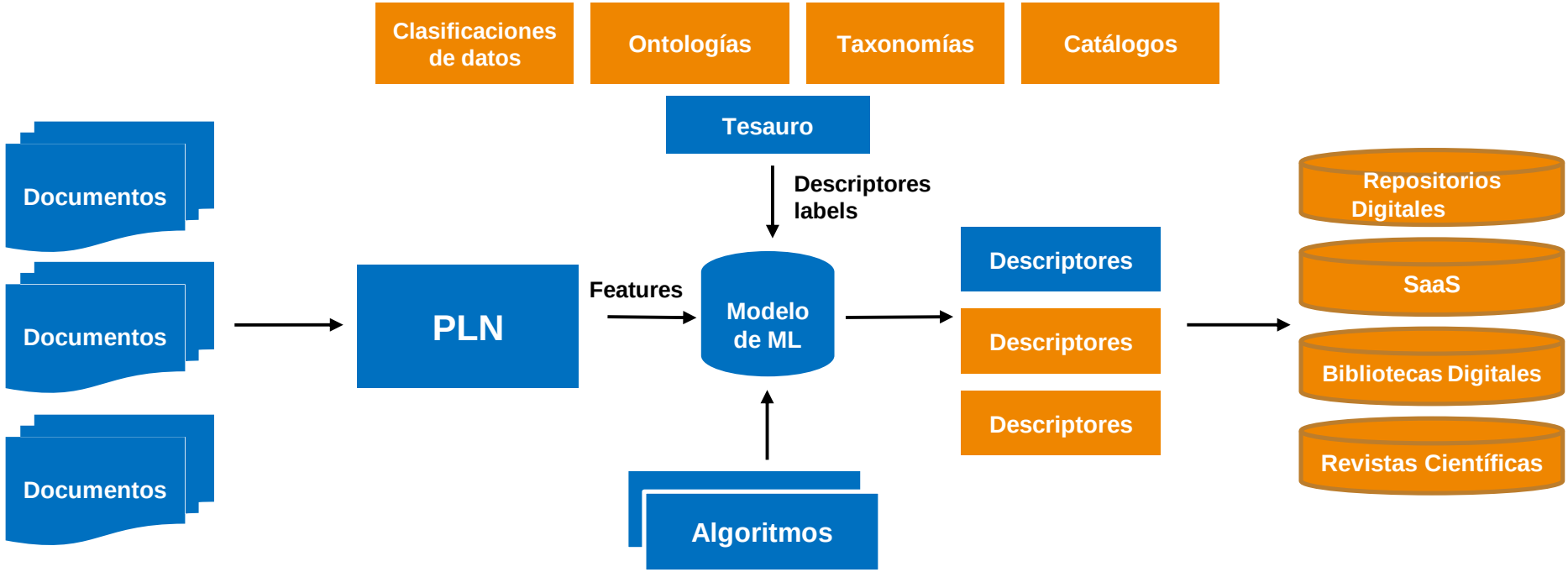
Lista de Encabezamientos de materia

Tesauros

BASE DE DATOS

RECUPERACIÓN DE INFORMACIÓN

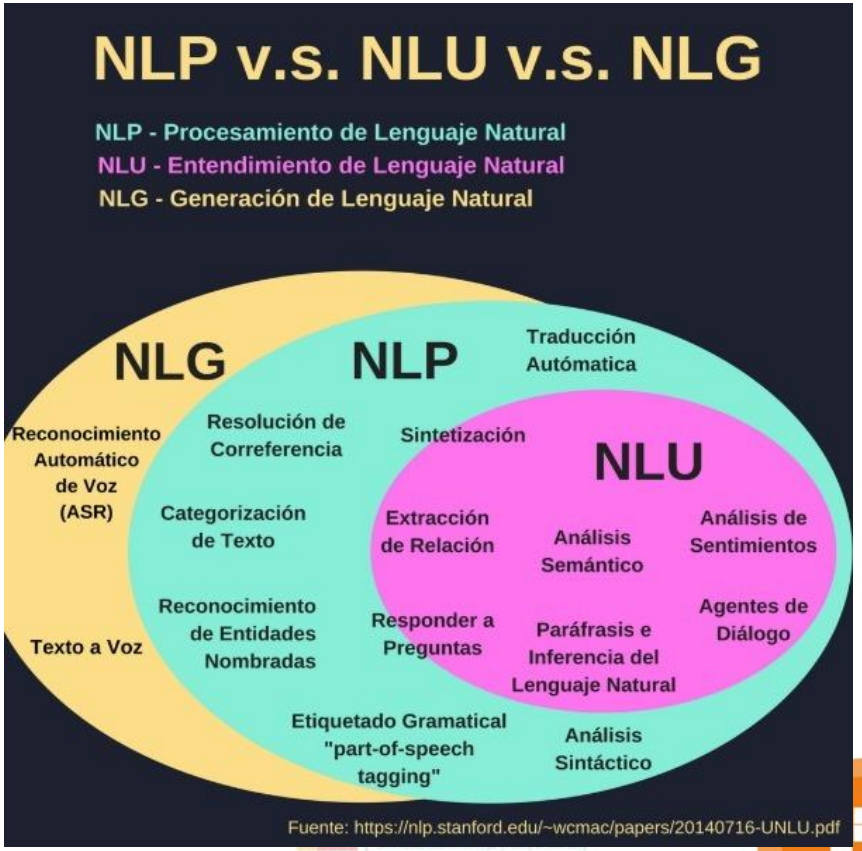
Proceso de la Indización basada en Inteligencia Artificial



Procesamiento del Lenguaje Natural (PLN)

Es un campo interdisciplinario que incorpora conocimientos de las ciencias computacionales, la IA y la lingüística computacional.

El NPL tiene como propósito lograr que la computadora entienda el lenguaje humano, hablado y/o escrito



Proceso del PLN

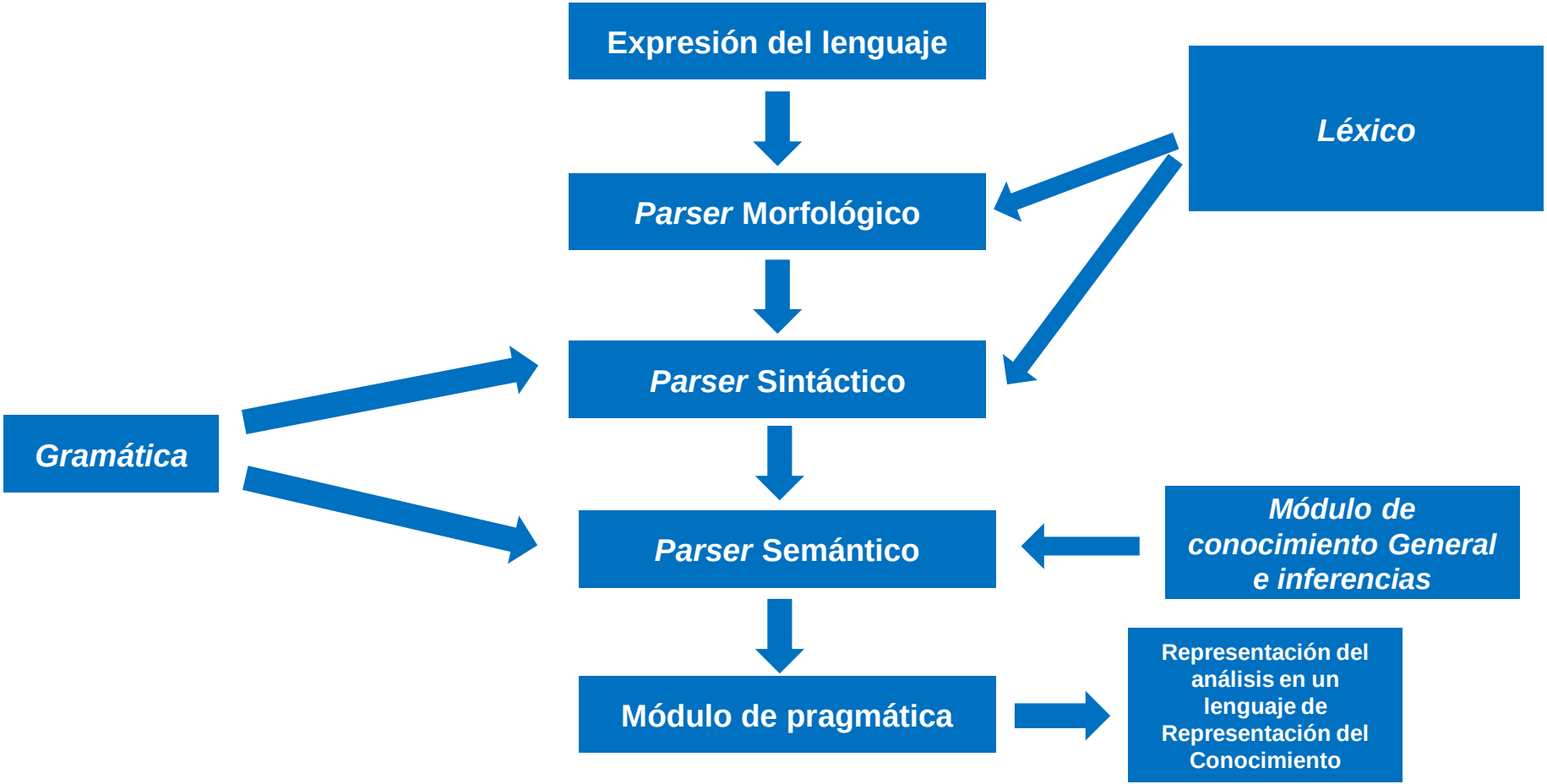
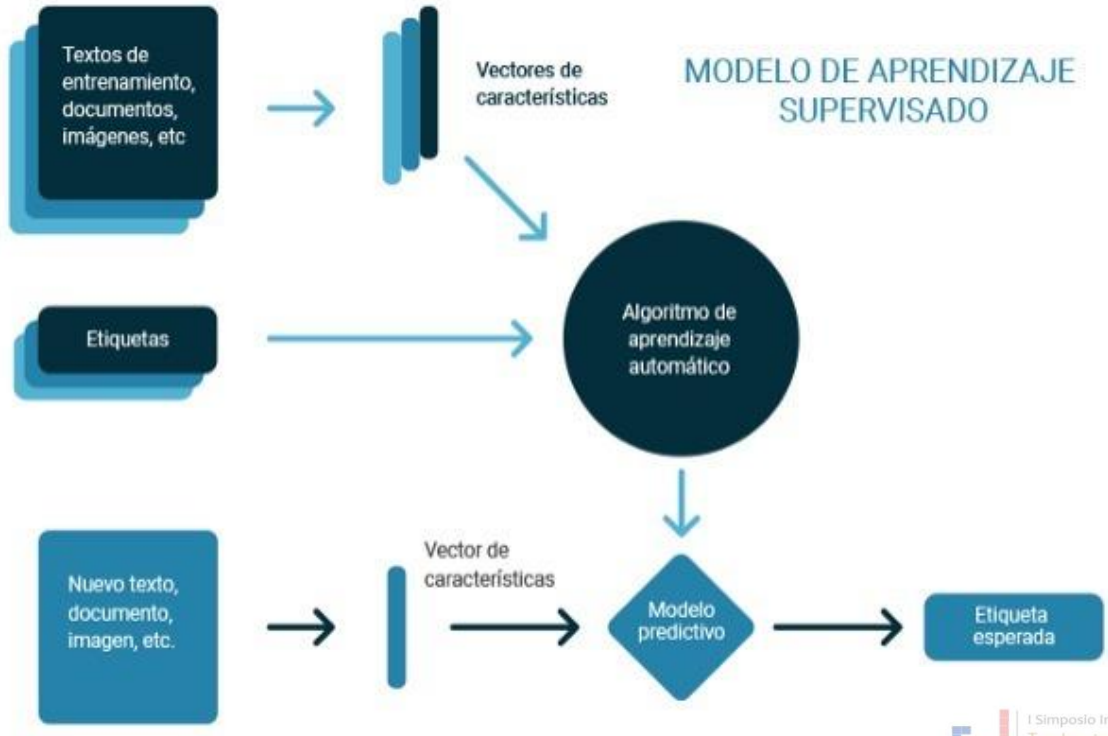
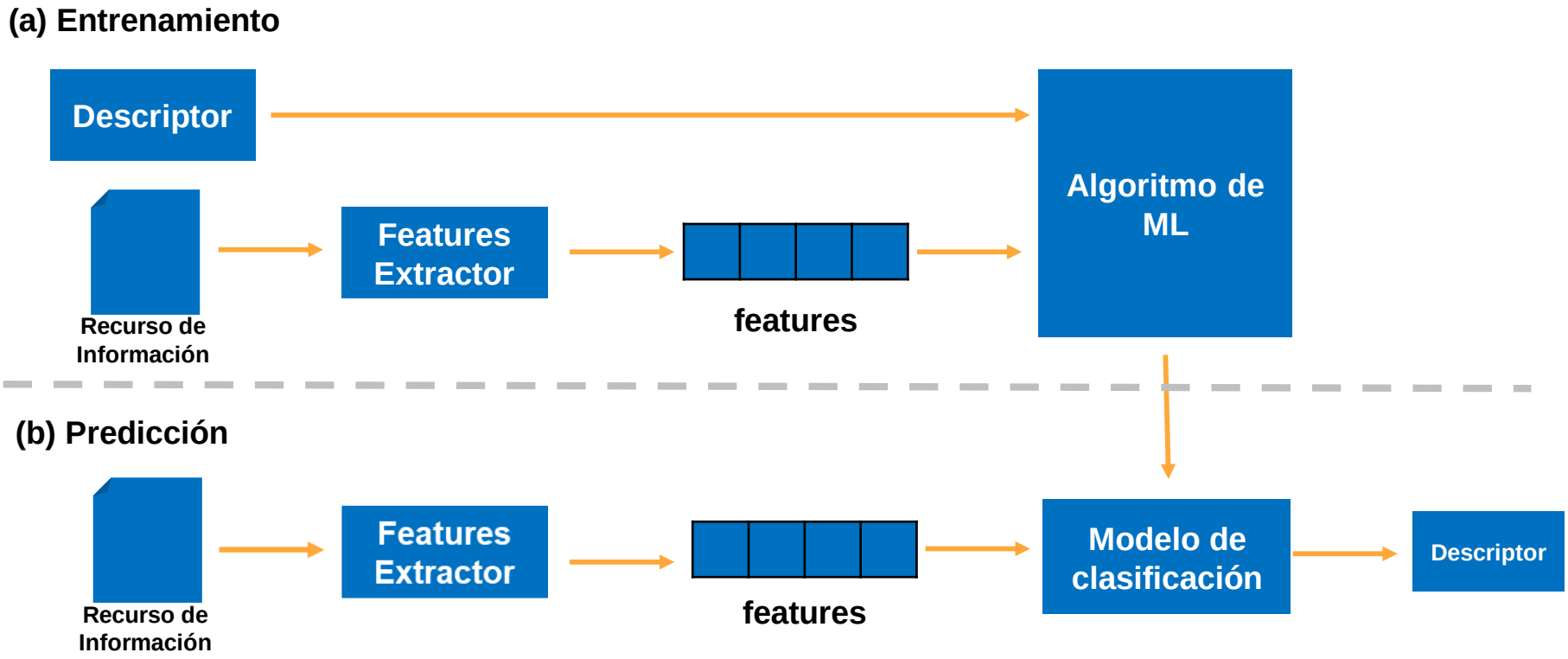


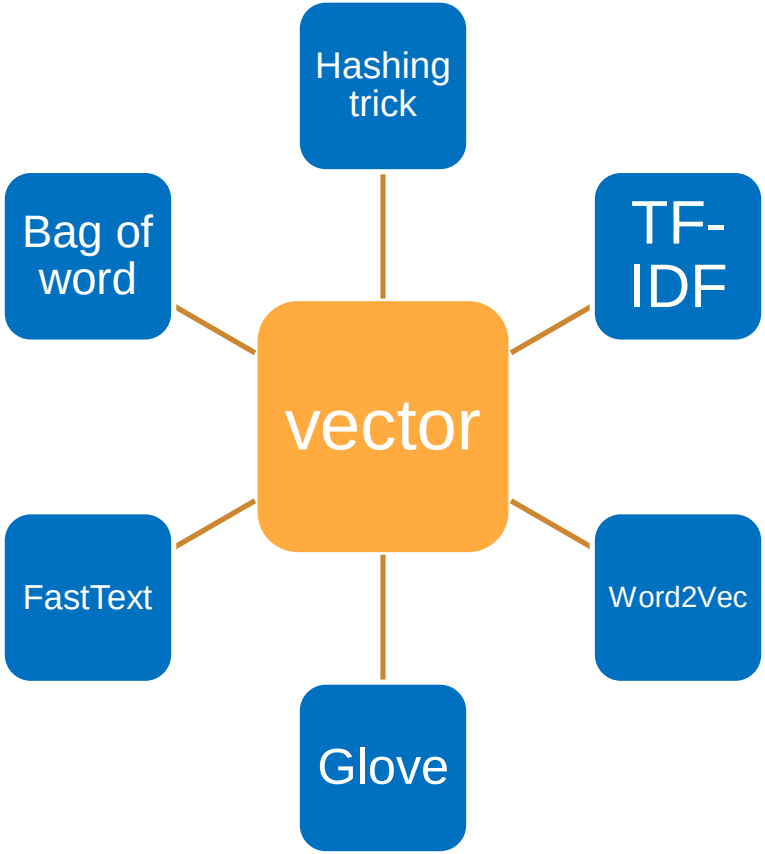
Diagrama de Flujo del Aprendizaje Supervisado



Aprendizaje supervisado: predicción



TIPOS DE VECTORIZACIÓN



Representación de “Bag of Words”

I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!

it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1

Bag of Words: vector de frecuencias

FEATURES: Propiedades o características del texto procesado.

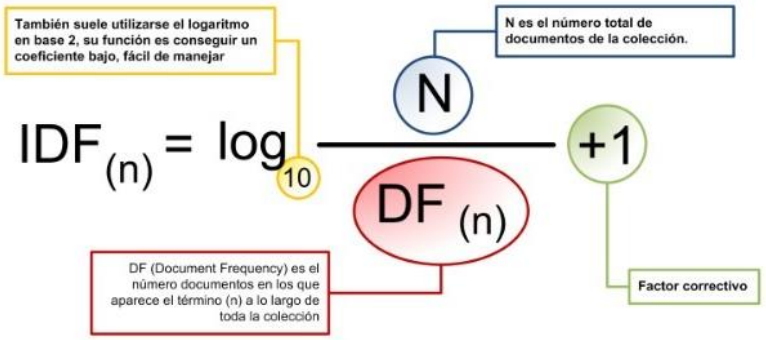
- **Pueden ser numéricas.**
- **Constituyen las entradas o insumos para el modelo.**
- **No redundantes.**
- **Facilitan el aprendizaje automático y la generalización.**
- **Es parte del proceso de reducción de la dimensionalidad.**

1. Factor TF: Factor de Frecuencia de Aparición de un Término

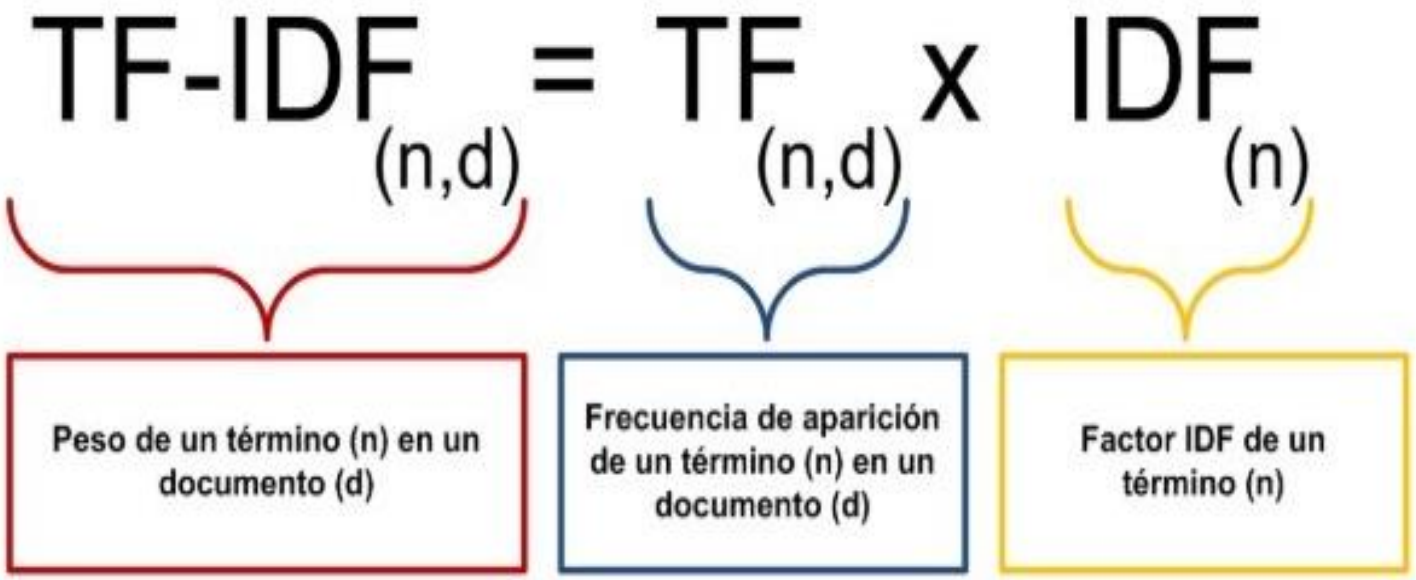
$$tf(n) = \sum_{(n)} D1$$

La frecuencia de aparición de un término (n) en un documento (D1) es la suma de las ocurrencias de dicho término

2. Factor IDF: Frecuencia Inversa del Documento para un Término



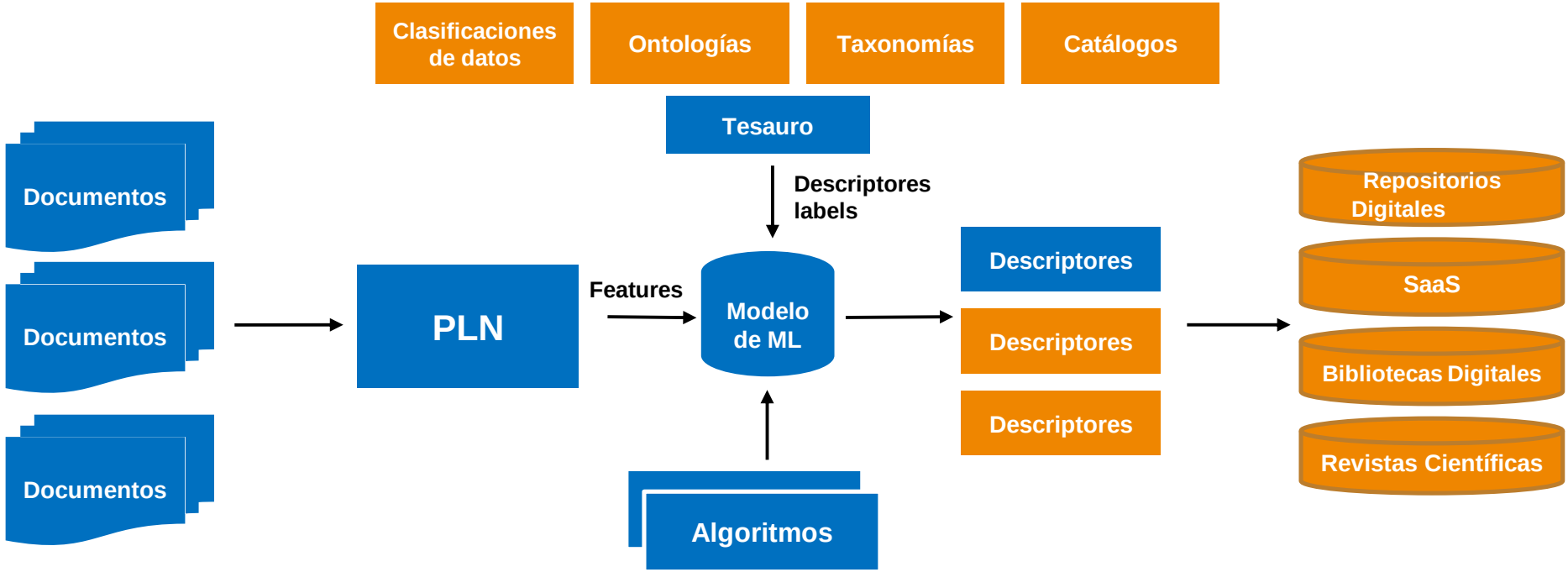
Peso TF-IDF para un término



Ejemplo de calculo de valores TF-IDF

Frecuencia de aparición de los términos TF			
Término	Doc1	Doc2	Doc3
biblioteca	27	4	24
archivo	3	33	0
documento	14	0	17
museo	0	33	29
Cálculo de Pesos TF-IDF			
biblioteca	TF-IDF(biblioteca,Doc1)	TF-IDF(biblioteca,Doc2)	TF-IDF(biblioteca,Doc3)
	27 x 2,65 = 71,55	4 x 2,65 = 10,60	24 x 2,65 = 63,60
archivo	TF-IDF(archivo,Doc1)	TF-IDF(archivo,Doc2)	TF-IDF(archivo,Doc3)
	3 x 3,08 = 9,24	33 x 3,08 = 101,64	0 x 3,08 = 0
documento	TF-IDF(documento,Doc1)	TF-IDF(documento,Doc2)	TF-IDF(documento,Doc3)
	14 x 2,50 = 35	0 x 2,50 = 0	17 x 2,50 = 42,50
museo	TF-IDF(museo,Doc1)	TF-IDF(museo,Doc2)	TF-IDF(museo,Doc3)
	0 x 2,62 = 0	33 x 2,62 = 86,46	29 x 2,62 = 75,98

Proceso de la Indización basada en Inteligencia Artificial



DEMOSTRACIÓN

I believe that good automatic indexer will exist once there's good artificial intelligence, something that presently doesn't exist.

Seth A. Maislin 2002

[Creo que existirán buenos índizadores automáticos una vez que haya buena inteligencia artificial, algo que actualmente no existe] Seth A. Maislin 2002