

Aspectos Generales

Marcelo Errecalde^{1,2}

¹Universidad Nacional de San Luis, Argentina 

²Universidad Nacional de la Patagonia Austral, Argentina 



Curso: Minería de Textos

Facultad de Informática - Universidad Nacional de La Plata
23 al 27 de Septiembre de 2019

Resumen

1

Introducción

- Lenguaje, lingüística y computación
- Minería de Texto (en contexto)
- Importancia del AAT
- Dificultades del PLN
- Tareas y Aplicaciones

2

El Proceso de KDD

3

KDD a partir de textos

4

Niveles del Lenguaje Natural

Introducción

- El lenguaje revela muchos aspectos de una persona ...



Introducción

- El lenguaje revela muchos aspectos de una persona ...



Introducción

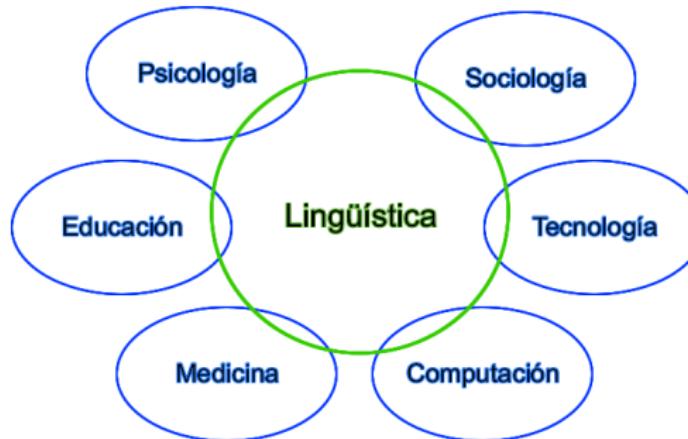
- El **lenguaje** revela muchos aspectos de una persona:
 - Pensamientos
 - Creencias
 - Sentimientos
 - Comportamientos
 - Personalidad
 - Nivel educativo
 - Orientación política
 - Edad
 - Género
 - y muchos más
 - Las **ciencias sociales** tienen una larga tradición en el estudio del **uso del lenguaje** para entender mejor estos aspectos.

Lingüística

Lingüística

Ciencia que estudia el lenguaje humano.

La Lingüística tiene intersección con distintos campos:



Lingüística y Computación

La intersección de ambas ciencias, da origen a distintas ramas que se denominan de diversas formas

Lingüística ∩ Computación

- Lingüística Computacional (**LC**).
- Procesamiento (Automático) del Lenguaje Natural (**PLN**).
 - Procesamiento/Análisis Automático de Textos (**PAT**).
- Tecnologías/Ingeniería del Lenguaje(**TL/IL**).
- Minería de Textos (**MT**)
- Otros

Lingüística y Computación

Lingüística Computacional (LC)

Se orienta a la construcción de modelos de lenguajes “entendibles” para las computadoras, es decir, **más formales** que los orientados al lector común.

Procesamiento (Automático) del Lenguaje Natural (PLN)

- Se ocupa más de los **aspectos técnicos, algorítmicos y matemáticos** de la aplicación de dichos modelos a grandes volúmenes de texto.
- Estos aspectos tienen que ver con la estructuración, extracción, y transformación de la información contenida en los textos.

Ambas disciplinas tienen el mismo objeto de estudio (el lenguaje natural), considerándolo desde enfoques diferentes.

Minería de Textos - Análisis Automático de Textos

En este curso nos centraremos en el enfoque usualmente denominado mediante alguno de los siguientes nombres:

- minería de textos (**MT**, *text mining*),
- análisis (automático) de textos (**AAT**, *text analytics*)
- o aprendizaje automático con textos (**machine learning from text**)

que podemos (intentar) definir como:

... el **proceso** de analizar **texto no estructurado** mediante técnicas de **aprendizaje automático** para extraer **conocimiento** (modelos **descriptivos** y **predictivos**) útil para propósitos particulares.

Minería de Texto - Análisis Automático de Textos

En este curso nos centraremos en el enfoque usualmente denominado mediante alguno de los siguientes nombres:

- minería de textos (**MT**, *text mining*),
- análisis (automático) de textos (**AAT**, *text analytics*)
- o aprendizaje automático con textos (**machine learning from text**)

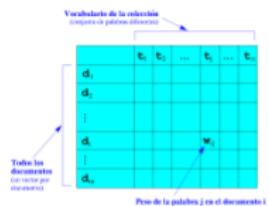
... y que surge de la interacción de **3 áreas** principales:

- Procesamiento del Lenguaje Natural (PLN)
- Recuperación de la Información (RI)
- Aprendizaje Automático (AA)

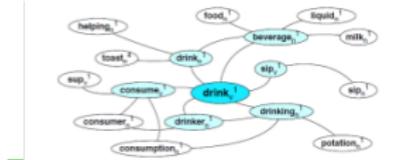
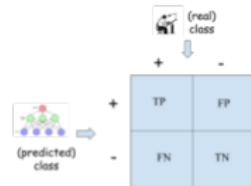
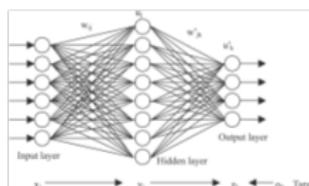
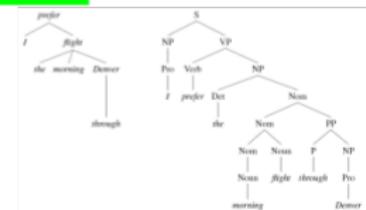
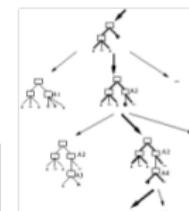
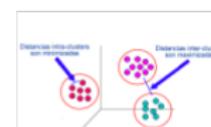
Minería de Texto en contexto



Minería de Texto (en contexto)

Minería de Texto en contexto

$$\begin{bmatrix} X \end{bmatrix} = \begin{bmatrix} V \end{bmatrix} \times \begin{bmatrix} W \end{bmatrix} = \begin{bmatrix} \alpha_1 & 0 & 0 & \dots & 0 \\ 0 & \alpha_2 & 0 & \dots & 0 \\ 0 & 0 & \alpha_3 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \alpha_n \end{bmatrix} \times \begin{bmatrix} m \times m \end{bmatrix} = \begin{bmatrix} m \times c \end{bmatrix}$$

**Recuperación de la Información****Procesamiento del Lenguaje Natural****Aprendizaje Automático**

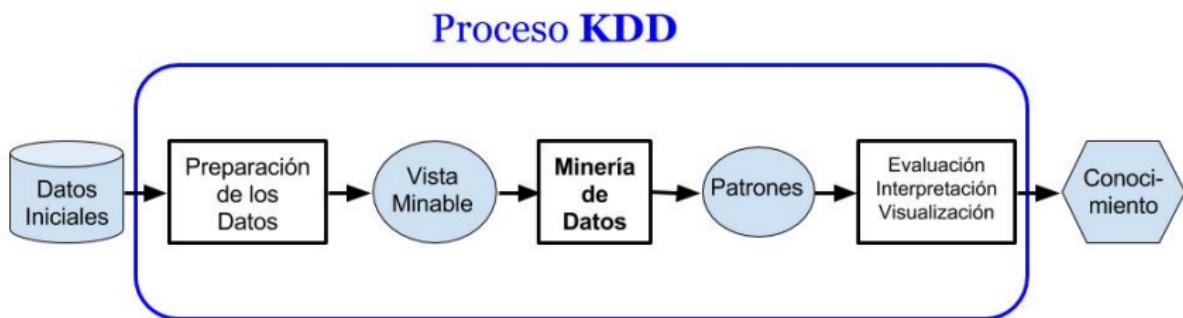
Minería de Texto en contexto

Estas 3 áreas, se relacionan directamente a dos diferentes concepciones respecto a cómo **visualizar** un texto y como representarlo:

- Textos como **bolsas de palabras**:
 - ➊ Originado en **RI**.
 - ➋ El orden de las palabras no es relevante.
 - ➌ Efectiva en aplicaciones como **clasificación, modelización de tópicos** y **sistemas recomendadores**.
- Textos como **conjuntos de secuencias**
 - ➊ Las sentencias individuales son extraídas como cadenas/secuencias.
 - ➋ Efectiva en aplicaciones que requieren mayor **interpretación semántica** (respuesta a preguntas (question answering), extracción de información, etc.).
 - ➌ Enfoque cercano a la **modelización del lenguaje** y el **PLN**.

Minería de Datos, de Textos y de la Web

Es útil considerar el proceso completo de **descubrimiento de conocimiento en Bases de Datos** (en inglés **Knowledge Discovery in Databases - KDD**)



Minería de Datos, de Textos y de la Web

Minería de Datos (MD)

- proceso de descubrir información útil (patrones), de forma automática, en grandes repositorios de datos.
- Incluye técnicas **estadísticas** y de **aprendizaje automático** para la extracción de **conocimiento**.

Minería de Textos (MT)

- **Preparación** de los datos: uso intensivo de **PLN/PAT/IR**.
- Tareas de MD + propias (**extracción de información**).

Minería de la Web (MW)

- ① Muchas similitudes con la MT.
- ② Agrega tareas propias de la Web (Minería de **estructura** y de **uso**).

Minería de Textos y Minería de Datos

Similitudes

Ambas dependen de

- Rutinas de pre-procesamiento.
- Algoritmos de descubrimiento de patrones.
- Elementos de la capa de presentación.

Diferencias

	Minería de Datos	Minería de Textos
Preprocesamiento	Asume datos estructurados	identificación y extracción de características representativas de documentos
Información	implícita	explícita
PLN	no relevante	clave

Minería de la Web

Minería de la *estructura* de la Web

- Descubre conocimiento útil a partir de los **(hiper)links**, los cuales representan la **estructura** de la Web.
- Ej.: páginas importantes, comunidades de usuarios, etc.

Minería del *contenido* de la Web

- Descubre conocimiento/información útil directamente del **contenido** de páginas Web.
- Incluye la categorización por **tópicos/temas**, pero también por **sentimientos, edad, sexo, nacionalidad, autoría**, etc.

Minería del *uso* de la Web

- Descubre **patrones de acceso** usando los **clicks** de los usuarios.

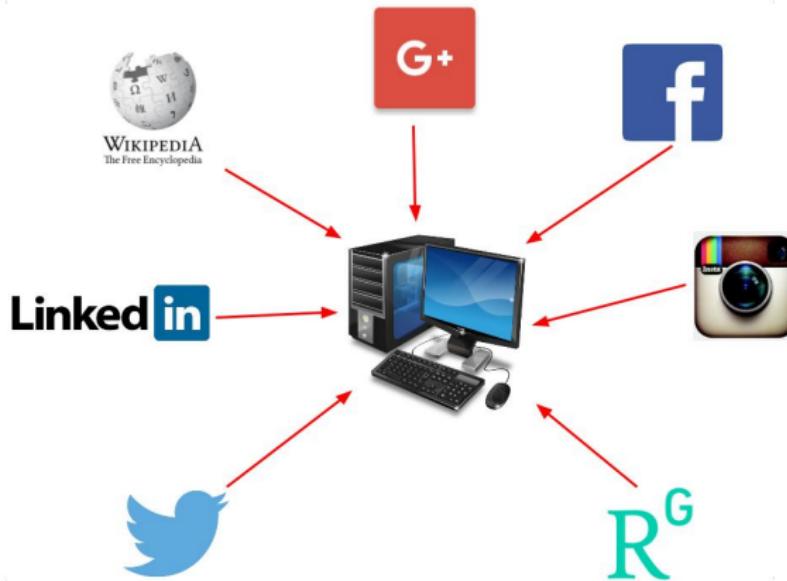
¿Porqué es *importante* el AAT?

- ① **Información y Conocimiento** es el recurso más importante que posee la raza humana.
- ② Este conocimiento se ha **comunicado, guardado y manejado** en la forma de **lenguaje natural** -griego, latín, inglés, español, etc
- ③ En la actualidad, además de almacenar este conocimiento en **libros y documentos**, también se lo hace en forma **digital**.

Importancia del AAT

¿Porqué es *importante* el AAT?

- ① “Inundación” de información textual (principal/ proveniente de la Web)



¿Porqué es *importante* el AAT?

- ① “Inundación” de información textual (principal/ proveniente de la Web)
- ② Importancia de inf. textual en la empresa (ley del 80%).
- ③ Texto: vehículo principal para el intercambio de información formal.
- ④ Texto: vehículo principal para el intercambio de inf. informal.
- ⑤ Aplicaciones de alto impacto: detección de spams, filtrado de noticias, detección de plagios, determinación del perfil del autor, análisis de opinión, análisis de tendencias, organización de patentes en categorías, clasificación y organización de páginas Web, etc.

Dificultades del PLN

- ① Principal dificultad del PLN: resolución de **ambigüedades** (acústicas, semánticas, etc)
- ② **Ambigüedad**: una unidad de lenguaje se puede interpretar de más de una manera:
 - a nivel de **palabra**: ejemplo, **fuerá**:
 - “como si fuera esta noche la última vez”, **ser**
 - “como si se fuera a la escuela”, **ir**
 - “está **fuerá** de la ciudad”, adverbio
 - a nivel de **oración**: “veo al gato con el telescopio”
- ③ Las ambigüedades se originan al considerar estas unidades fuera del **contexto local y global**.
- ④ Su resolución implica un análisis a un nivel mayor que considere 3 tipos de conocimiento:
 - **Conocimiento lingüístico**
 - **Conocimiento extra-lingüístico**
 - **Conocimiento obtenido del mismo texto**

Tareas “soporte” del AAT

- Segmentado y normalización de textos
- Etiquetados de componentes textuales
 - Etiquetado en categorías gramaticales (**POS tagging**)
 - Desambiguación del sentido de las palabras (**WSD**)
 - Reconocimiento de entidades nombradas (**NER**)
- Extracción de relaciones

Tareas “intermedias” del AAT

- Ayuda en la Preparación de Textos.
- Búsqueda de Información.
- Categorización temática
- Modelización de tópicos
- Extracción de información
- Análisis de autoría
 - ① Atribución de autoría (authorship attribution)
 - ② Detección de plagio (plagiarism detection)
 - ③ Determinación del perfil del autor (author profiling)
 - ④ Detección de inconsistencias estilísticas
- Análisis de sentimiento ? Análisis de sentimiento por aspectos?

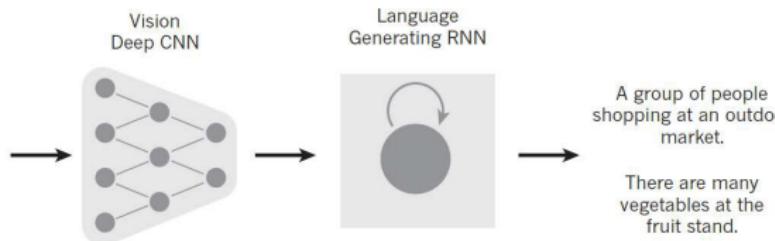
Tareas “avanzadas” del AAT

- ① Generación automática de **respuestas** (Question Answering).
- ② Generación automática de **resúmenes** (Summarization).
- ③ **Traducción** automática
- ④ **Interfaces** en Lenguaje Natural.
- ⑤ **Clasificación anticipada** de textos
- ⑥ Tareas **multi-modales** y **temporales** (question visual answering, scene description, early risk detection)

Aplicaciones

- 1 Propaganda dirigida
- 2 Evaluación de campañas, políticos y productos
- 3 Detección de **plagio** (intrínseco y extrínseco) - plagio de patentes
- 4 Detección de “bullying” y pedófilos en la Web
- 5 Detección anticipada de signos de **depresión, anorexia, rumores** y tendencias **suicidas**
- 6 Atribución de **autoría** en casos forenses
- 7 Determinación del **perfil** del autor (sexo, grupo etario, nacionalidad, rasgos de personalidad, orientación política, sexual y religiosa)
- 8 Identificación de **información de calidad**
- 9 **Predicción** del comportamiento del mercado, precios de productos y servicios, ganancias obtenidas en un rubro, difusión de enfermedades y ocurrencia de catástrofes.

Aplicaciones recientes: descripción de escenas



A group of people
shopping at an outdoor
market.

There are many
vegetables at the
fruit stand.

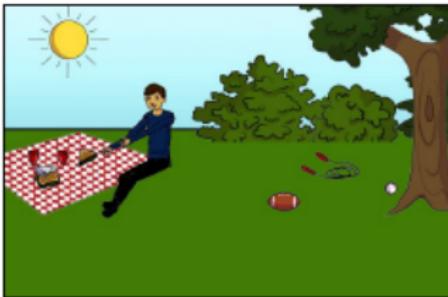
Aplicaciones recientes: responder preguntas sobre imágenes



What color are her eyes?
What is the mustache made of?



How many slices of pizza are there?
Is this a vegetarian pizza?



Is this person expecting company?
What is just under the tree?



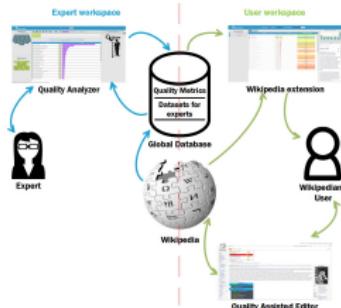
Does it appear to be rainy?
Does this person have 20/20 vision?

Aplicaciones recientes: calidad de información en Wikipedia

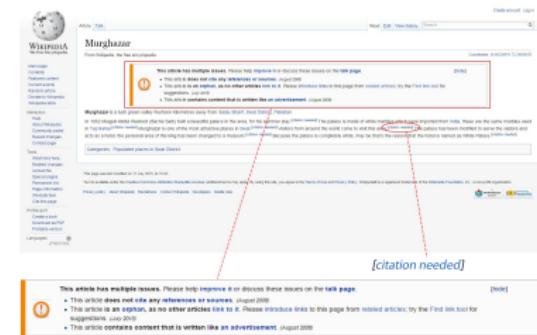
Artículos Destacados



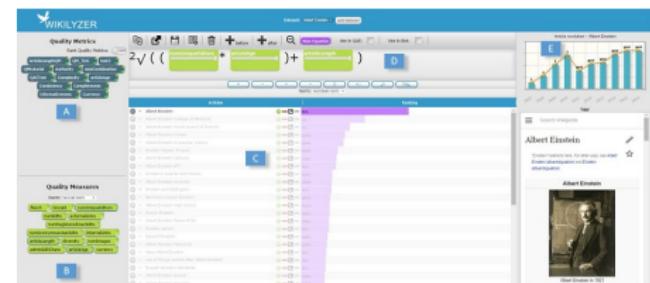
Visualización



Detección de Fallas



Métricas de Calidad



Tareas y Aplicaciones

Perfil de autor



Tareas: determinar

- ① edad
- ② género
- ③ personalidad
- ④ orientación política
- ⑤ ...

Detección de pedófilos en la Web

- Datos de entrenamiento en
www.perverted-justice.com
- Competencias recientes
pan.webis.de/clef12/pan12-web

Ejemplo:

Example 1: what nationality are u?

Exchange of information

Example 2: what r u wearing?

Grooming

Example 3: would u let me?

Example 4: thing it is me feeling u

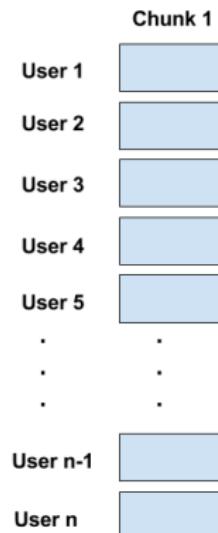
Example 5: what's your address?

Example 6: can I stay at your house overnight if i go?

Approach

Clasificación temprana / detección temprana de riesgos

- Entrenar con información secuencial **completa**.
- Luego clasificar, **tan pronto como sea posible**



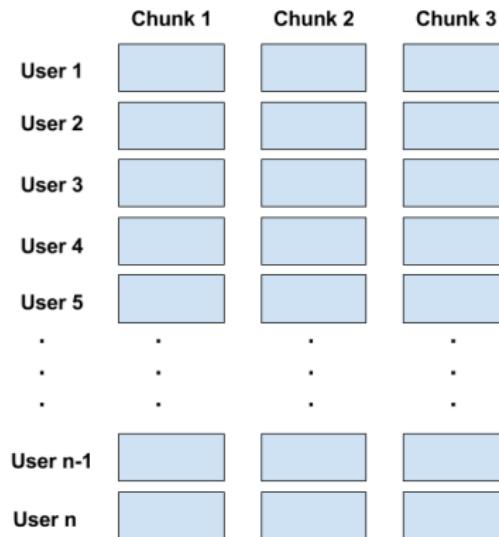
Clasificación temprana / detección temprana de riesgos

- Entrenar con información secuencial **completa**.
- Luego clasificar, **tan pronto como sea posible**

	Chunk 1	Chunk 2
User 1		
User 2		
User 3		
User 4		
User 5		
⋮	⋮	⋮
User n-1		
User n		

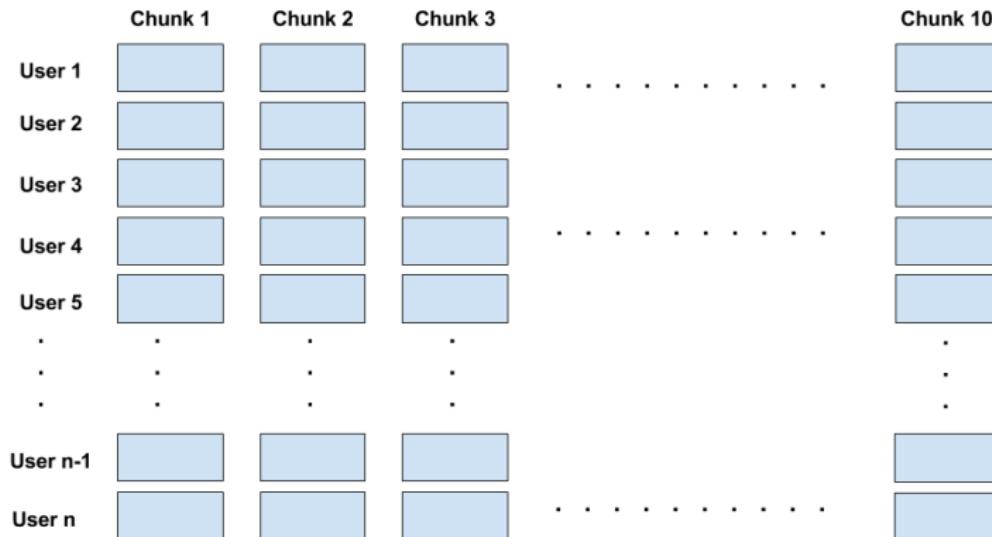
Clasificación temprana / detección temprana de riesgos

- Entrenar con información secuencial **completa**.
- Luego clasificar, **tan pronto como sea posible**



Clasificación temprana / detección temprana de riesgos

- Entrenar con información secuencial completa
 - Luego clasificar, tan pronto como sea posible



Aplicaciones recientes: early risk detection

A grayscale photograph of a person's profile, facing right, with their hair blowing in the wind.

ERISK 2019

ABOUT CALL FOR CONTRIBUTIONS IMPORTANT DATES SCHEDULE ORGANIZERS 2018 2017 CONTACT

ERISK 2019: EARLY RISK PREDICTION ON THE INTERNET

CLEF 2019 Workshop

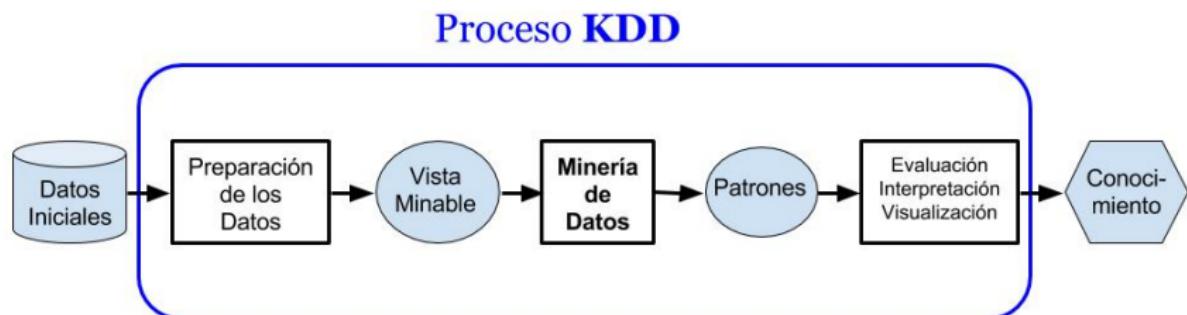
Lugano, 09–12 September 2019

FIND OUT MORE

Minería de Datos y KDD

KDD (Knowledge Discovery in Databases): Proceso no trivial de identificar **patrones** válidos, novedosos, potencialmente útiles y, en última instancia, comprensibles, a partir de los datos.

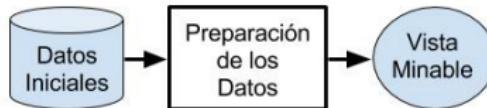
Fases del Proceso KDD



Fase de Preparación de los datos

KDD (*Knowledge Discovery in Databases*): Proceso no trivial de identificar **patrones** válidos, novedosos, potencialmente útiles y, en última instancia, comprensibles, a partir de los datos.

Fases del Proceso KDD



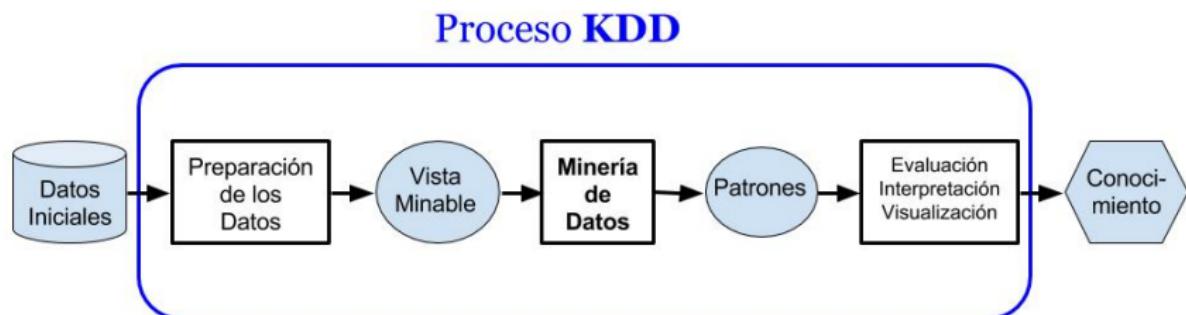
Fase de Preparación de los datos

- Sub-fase de **recopilación e integración de los datos**
 - Determinar fuentes de información útiles y dónde conseguirlas.
 - Coleccionar múltiples bases de datos **heterogéneas** en un único **almacén de datos**.
- Sub-fase de **selección, limpieza y transformación**
 - Detección de valores **anómalos** (no siempre eliminados).
 - Tratamiento de datos **faltantes** (o perdidos).
 - Selección de atributos **relevantes** (columnas).
 - Selección de una **muestra** de datos (filas).
 - Construcción de **nuevos atributos** (agrupamiento, numerización, discretización, normalización).

Minería de Datos y KDD

KDD (Knowledge Discovery in Databases): Proceso no trivial de identificar **patrones** válidos, novedosos, potencialmente útiles y, en última instancia, comprensibles, a partir de los datos.

Fases del Proceso KDD



Fase de Minería de datos

KDD (*Knowledge Discovery in Databases*): Proceso no trivial de identificar **patrones** válidos, novedosos, potencialmente útiles y, en última instancia, comprensibles, a partir de los datos.

Fases del Proceso KDD



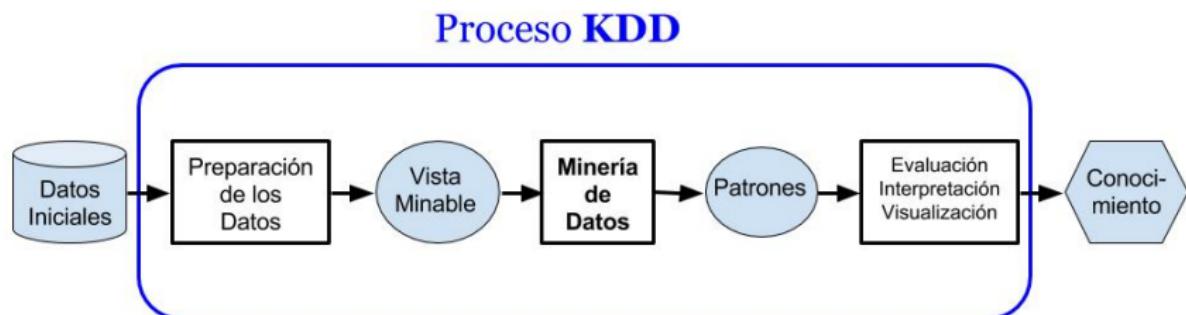
Fase de Minería de datos

- Determinar qué tipo de **tarea** de MD es el más apropiado (clasificación, agrupamiento, etc).
- Elegir tipo de **modelo** (árboles de decisión, reglas de clasificación, Redes Neuronales).
- Elegir el **algoritmo** de minería (o aprendizaje) (CART, C5.0, Backpropagation)

Minería de Datos y KDD

KDD (Knowledge Discovery in Databases): Proceso no trivial de identificar **patrones** válidos, novedosos, potencialmente útiles y, en última instancia, comprensibles, a partir de los datos.

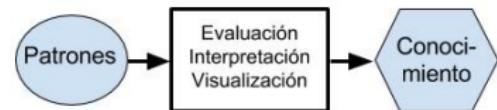
Fases del Proceso KDD



Fase de evaluación, interpretación y visualización

KDD (*Knowledge Discovery in Databases*): Proceso no trivial de identificar **patrones** válidos, novedosos, potencialmente útiles y, en última instancia, comprensibles, a partir de los datos.

Fases del Proceso KDD



Fase de evaluación, interpretación y visualización

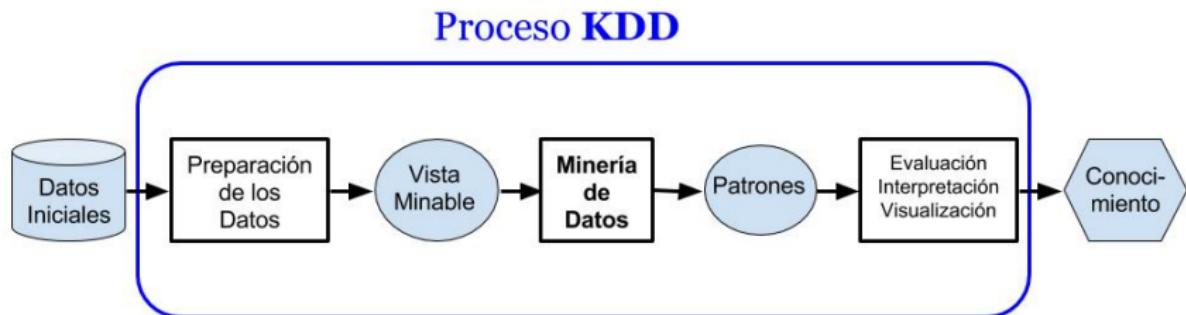
Criterios para la **evaluación** de los modelos (patrones) descubiertos:

- **Precisos**
- **Comprensibles** (inteligibles)
- **Interesantes** (útiles y novedosos)

KDD a partir de textos

Sirve de guía para la organización de los contenidos de este curso.

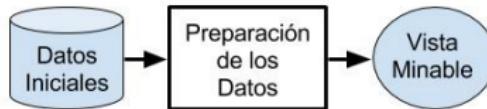
Fases del Proceso KDD



Fase de Preparación de los textos

Sirve de guía para la organización de los contenidos de este curso.

Fases del Proceso KDD



Fase de Preparación de los textos

- Recopilación de textos
- Pre-procesamiento de los textos (**clase 2**)
- Representación de los textos (**clase 3**)
- Reducción de dimensionalidad (**clase 4**)

Algunas técnicas de pre-procesamiento

- ➊ Partición del texto
- ➋ Filtrado (“stop-words”, baja frecuencia)
- ➌ Normalización (mayúsculas, variaciones de uso)
- ➍ Truncado (“stemming”) y lematización (“lemmatization”)
- ➎ Etiquetado de las palabras
 - De Partes de la Oración (Part of Speech (**POS**) Tagging)
 - Desambiguación del Significado de las Palabras (**WSD**)
 - Reconocimiento de Entidades Nombradas (**NER**)

Tokenización

Ejemplo, dada la siguiente sentencia:

After sleeping for four hours, he decided to sleep for another four.

El proceso de tokenización produciría:

{ `After` “sleeping” “for” “four” “hours” “he” “decided”
“to” “sleep” “for” “another” “four” }.

Comparación truncado vs lematización

In [1]:

```
compare_normalization(u"Our meeting today was worse than yesterday, "
                      "I'm scared of meeting the clients tomorrow.")
```

Out[1]:

Lemmatization:

```
['our', 'meeting', 'today', 'be', 'bad', 'than', 'yesterday', ',', ,
'i', 'be', 'scared', 'of', 'meet', 'the', 'client', 'tomorrow', '.']
```

Stemming:

```
['our', 'meet', 'today', 'wa', 'wors', 'than', 'yesterday', ',', ,
'i', "'m", 'scare', 'of', 'meet', 'the', 'client', 'tomorrow', '.']
```

Etiquetado de las Categorías Gramaticales

Ejemplo: las palabras de la sentencia

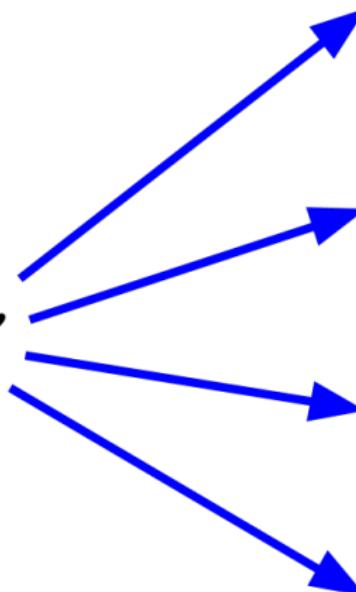
The grand jury commented on a number of other topics.

podrían ser etiquetadas con las siguientes categorías por un sistema de ECG:

The/**DT** grand/**JJ** jury/**NN** commented/**VBD** on/**IN** a/**DT** number/**NN** of/**IN** other/**JJ** topics/**NNS** ./.

¿Qué significa “banco”?

“banco”



Desambiguación del Significado de las Palabras

Definición

En inglés **Word Sense Disambiguation (WSD)** trata de resolver la ambigüedad en el significado de las palabras (o frases) en base al contexto en que éstas aparecen.

Ejemplo, la palabra banco...

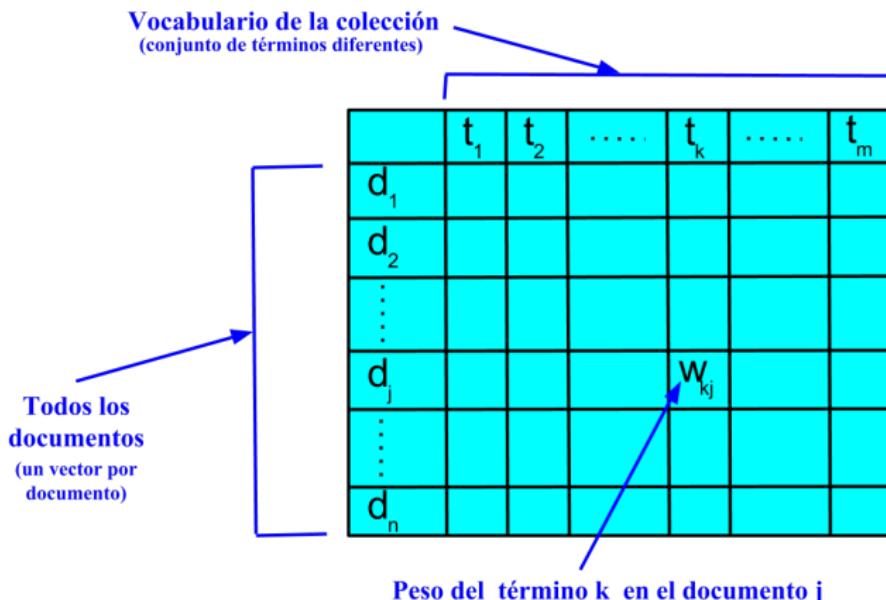
Frase		Significado
Perdí la mañana en el banco pagando impuestos.	⇒	Institución Financiera
Desde el barco vi el banco de peces.	⇒	Cardumen
Sentado en un banco suspiraba.	⇒	Mueble
Las donaciones se están recibiendo en el banco de sangre.	⇒	Establecimiento Médico

Reconocimiento de Entidades Nombradas

Características

- Subtarea de Extracción de Información.
- Consiste en la **ubicación** y **clasificación** de palabras dentro de un texto, en categorías tales como **nombres de personas, lugares, organizaciones, cantidades**, etc.
- El resultado de este proceso es un documento con etiquetas que identifican el comienzo y fin de las entidades nombradas.
- Ejemplo: “Jim bought 300 shares of Acme Corp. in 2006”
⇒ <ENAMEX TYPE=“PERSON”>Jim</ENAMEX> bought <NUMEX TYPE=“QUANTITY”>300</NUMEX> shares of <ENAMEX TYPE=“ORGANIZATION”>Acme Corp.</ENAMEX> in <TIMEX TYPE=“DATE”>2006</TIMEX>.

Representación vectorial de documentos: visión general



Representación de *bolsa de palabras* ("Bag of Words" (BoW))

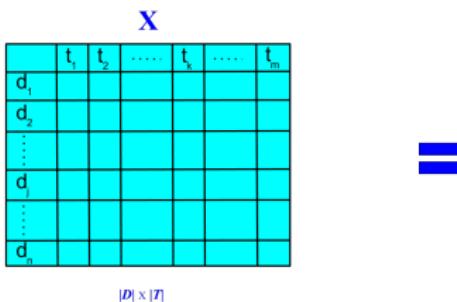
Documentos

- ① "pintaron el banco de la plaza"
- ② "si paso la prueba, iremos paso a paso"
- ③ "no me banco ir al banco a cobrar cheques"

Representación BoW

D/T	a	al	banco	cheques	cobrar	de	el	ir	iremos	la	me	no	paso	pintaron	plaza	prueba	si
d1	0	0	1	0	0	1	1	0	0	1	0	0	0	1	1	0	0
d2	1	0	0	0	0	0	0	0	1	1	0	0	3	0	0	1	1
d3	1	1	2	1	1	0	0	1	0	0	1	1	0	0	0	0	0

Reducción de dimensionalidad: LSA



=

$$X = U_k S_k V_k^T$$

U_k

	c_1	c_2	c_k
d_1				
d_2				
\vdots				
d_j				
\vdots				
d_n				

$|D| \times k$

S_k

	c_1	c_2	c_k
c_1				
c_2				
\vdots				
c_k				

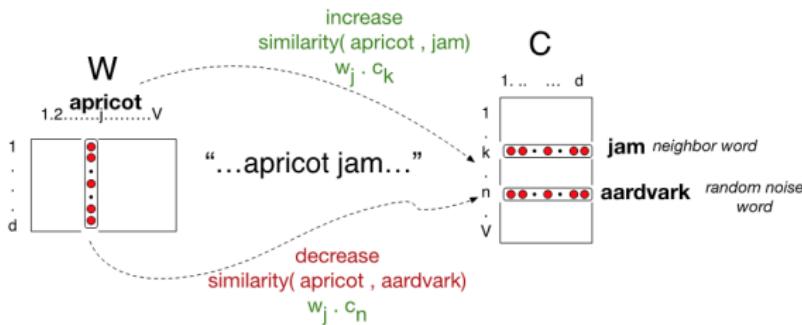
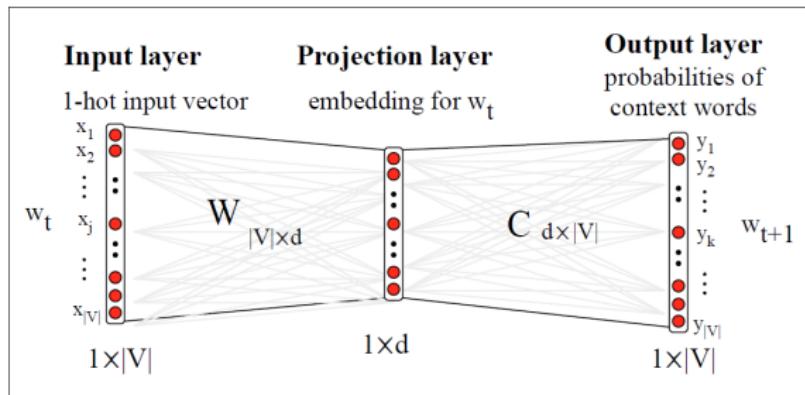
$k \times k$

V_k^T

	t_1	t_2	t_i	t_m
c_1						
c_2						
\vdots						
c_k						

$k \times |T|$

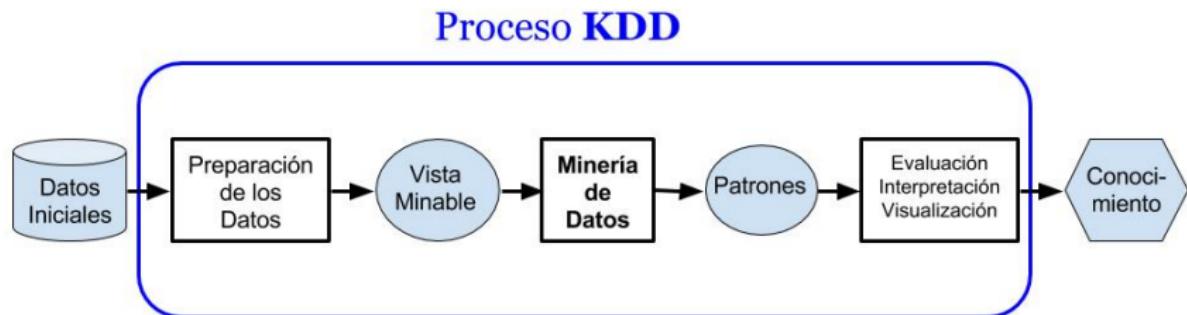
Reducción de dimensionalidad: word2vec



KDD a partir de textos

Sirve de guía para la organización de los contenidos de este curso.

Fases del Proceso KDD



Fase de Minería de textos

Sirve de guía para la organización de los contenidos de este curso.

Fases del Proceso KDD



Fase de Minería de textos

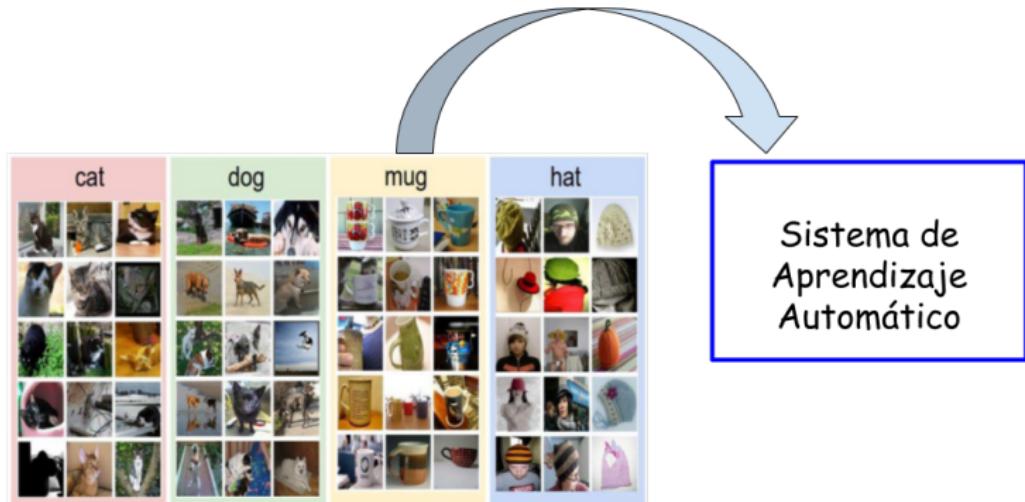
- Categorización de textos (clase 5 - Parte A)
 - Palabras
 - Documentos completos
- Agrupamiento de textos (clase 5 - Parte B)
 - Palabras
 - Documentos completos
- Extracción de Información
 - Cerrada
 - Abierta

Repaso: tipos de aprendizaje automático

- **Aprendizaje supervisado:** experiencia es un conjunto de ejemplos $\langle x, f(x) \rangle$, de la función f a ser aproximada.
- **Aprendizaje por refuerzos:** experiencia son secuencias de tri-uplas $\langle s, a, r \rangle$, donde a es la acción tomada por el agente en el estado s , y r es la evaluación numérica recibida desde el ambiente por la realización de esta acción.
- **Aprendizaje no supervisado:** **no existe** una retroalimentación explícita desde el ambiente.

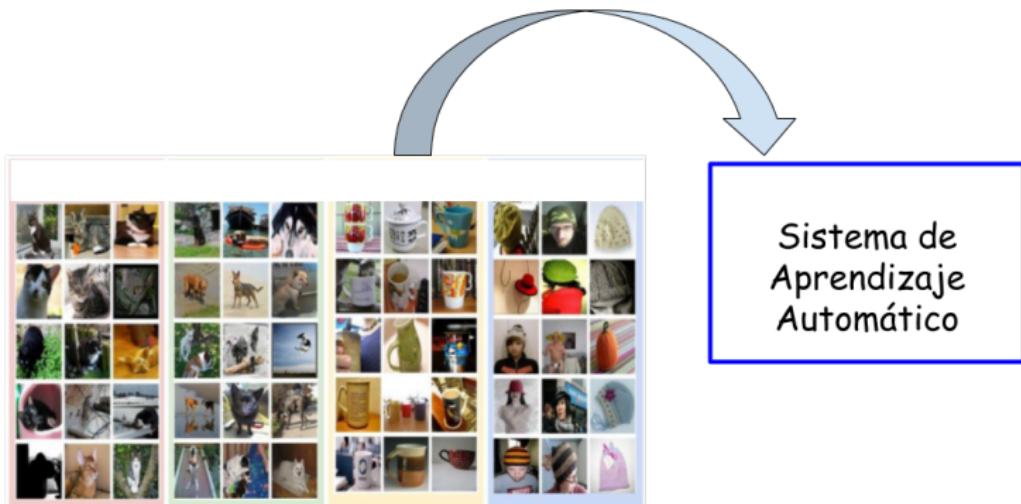
Retroalimentación en Aprendizaje Automático: supervisado

Retroalimentación = Ejemplos Etiquetados



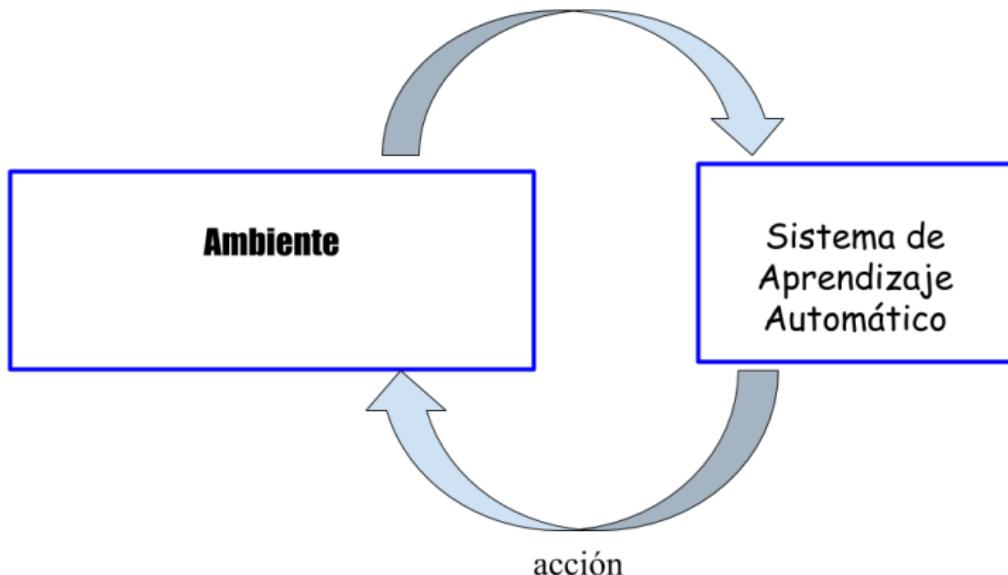
Retroalimentación en Aprendizaje Automático: no-supervisado

Retroalimentación = Ejemplos No Etiquetados



Retroalimentación en Aprendizaje Automático: refuerzo

Retroalimentación = Recompensa/penalización



Categorización de textos

Datos

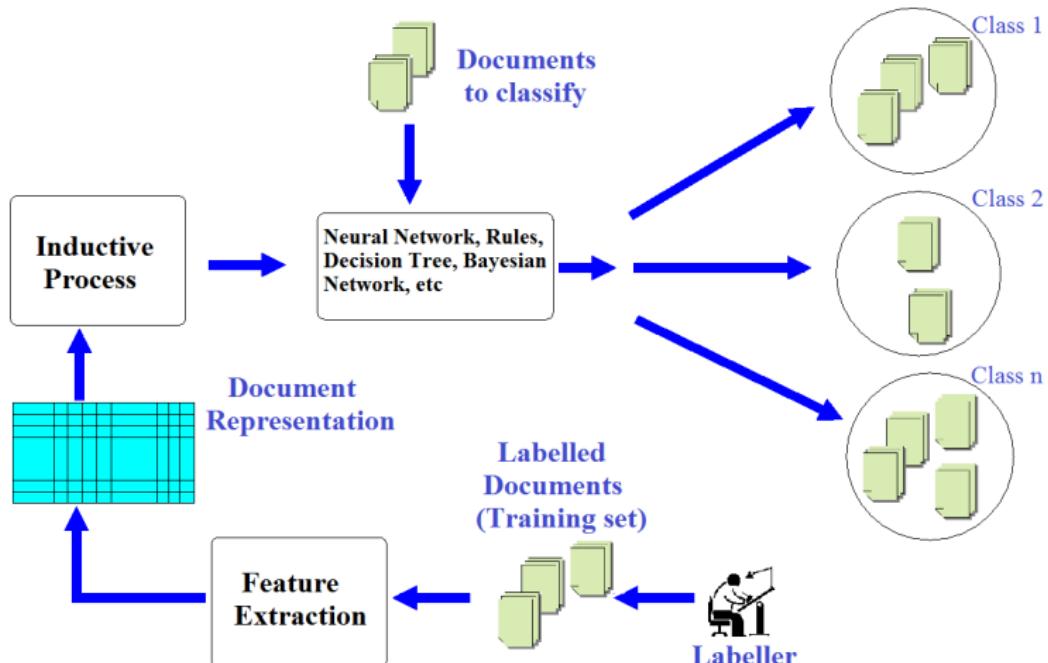
- Una colección de documentos \mathcal{D}
- Un conjunto de categorías $\mathcal{C} = \{c_1, \dots, c_{|\mathcal{C}|}\}$

Categorización de textos es la tarea de asignar los documentos en \mathcal{D} a las categorías en \mathcal{C}

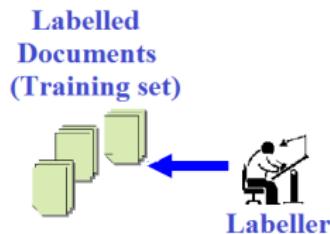
Ejemplos:

Problema	Objeto	Categorías (\mathcal{C})
Detección de spam	e-mails	$\{\text{True}, \text{False}\}$
Identificación de autores	documentos	autores
Categorización de noticias	cables de noticias	secciones del periódico
WSD	palabras con su contexto	significados de la palabra

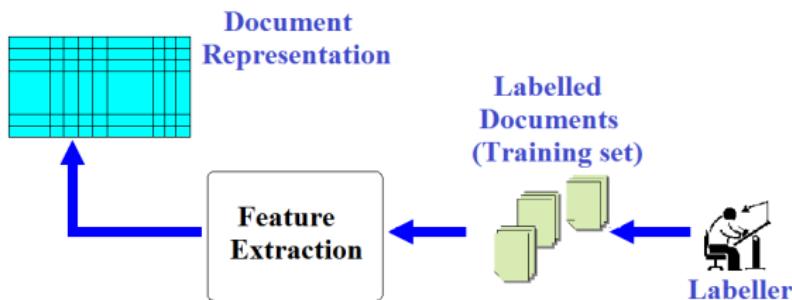
Ejemplo: Categorización Supervisada de Textos



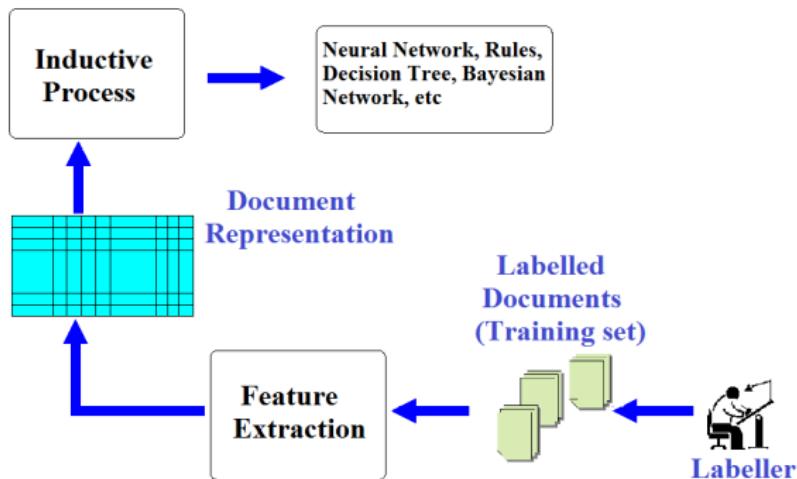
Ejemplo: Categorización Supervisada de Textos



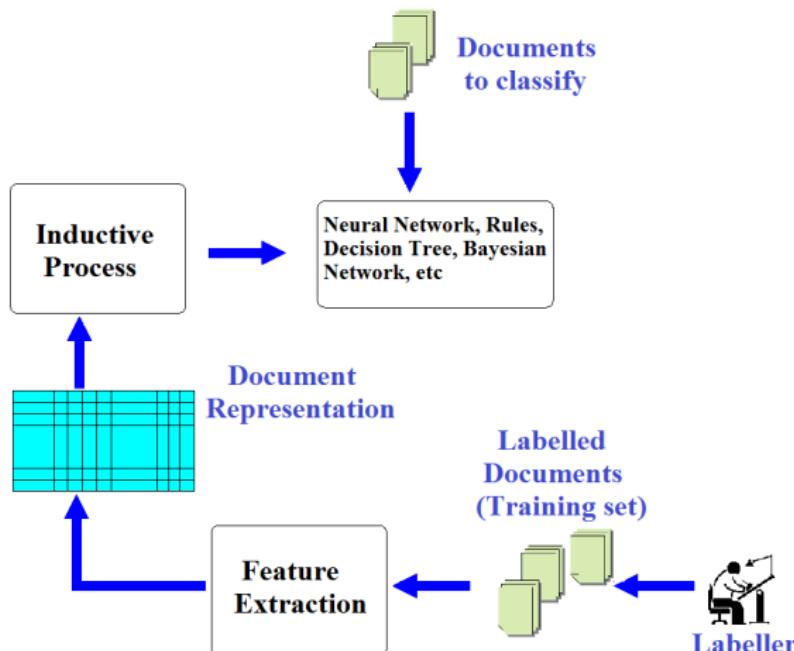
Ejemplo: Categorización Supervisada de Textos



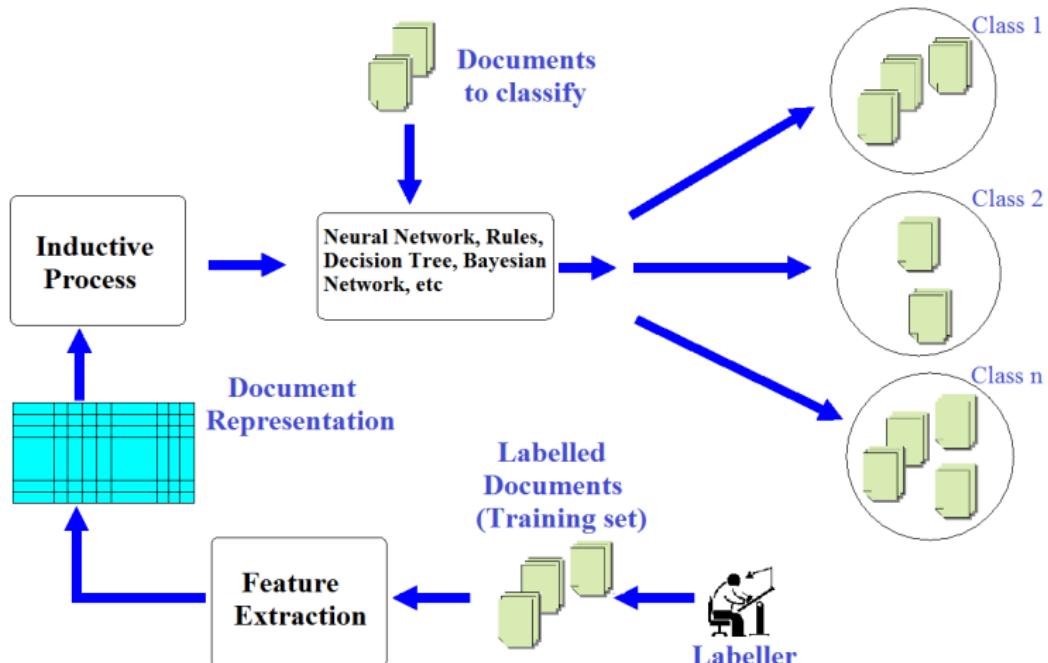
Ejemplo: Categorización Supervisada de Textos



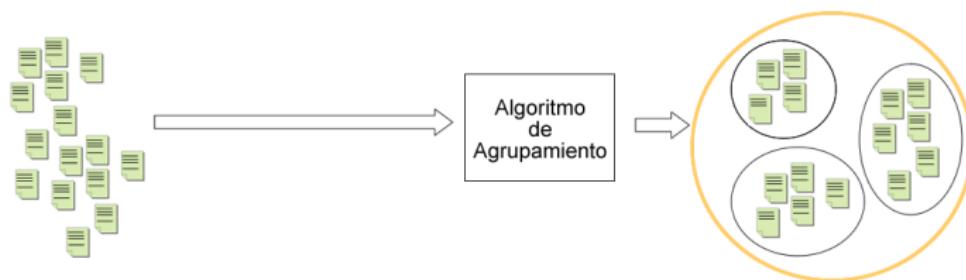
Ejemplo: Categorización Supervisada de Textos



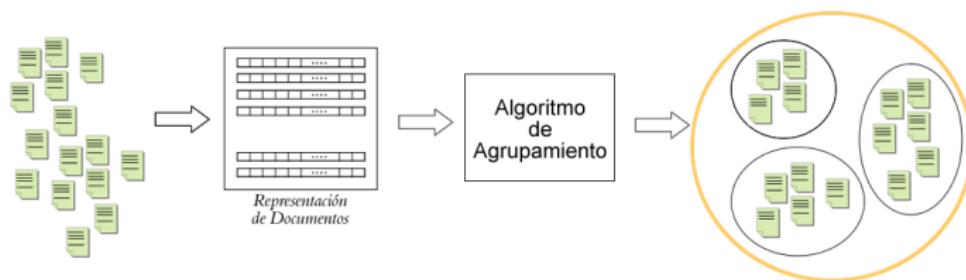
Ejemplo: Categorización Supervisada de Textos



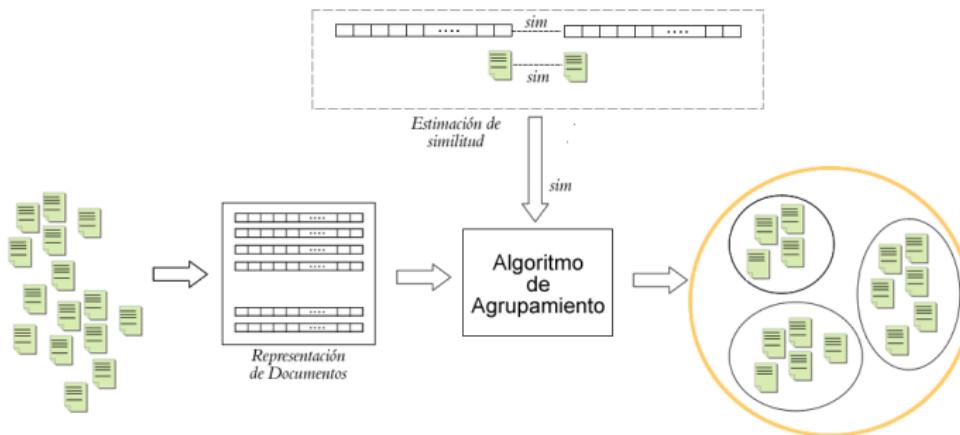
Agrupamiento de documentos



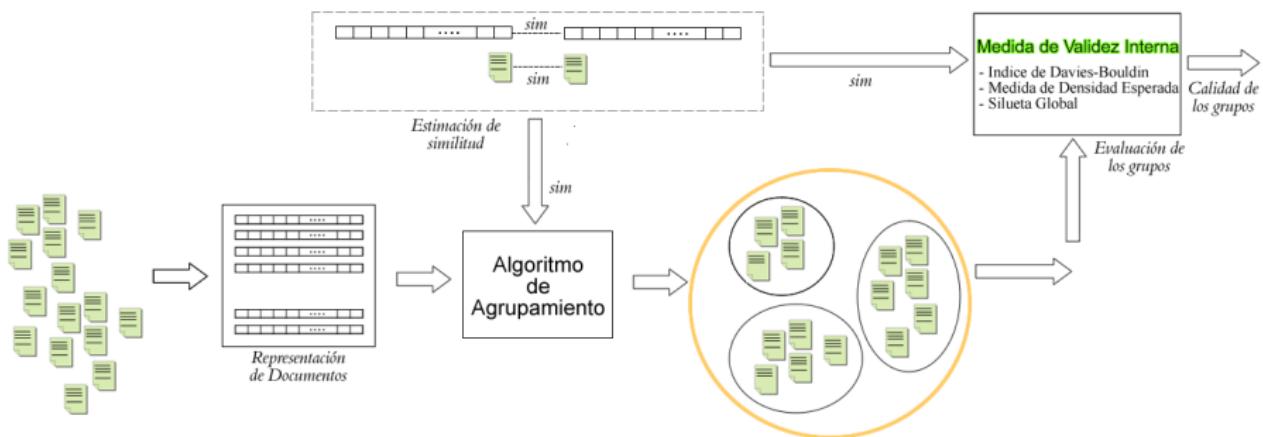
Agrupamiento de documentos



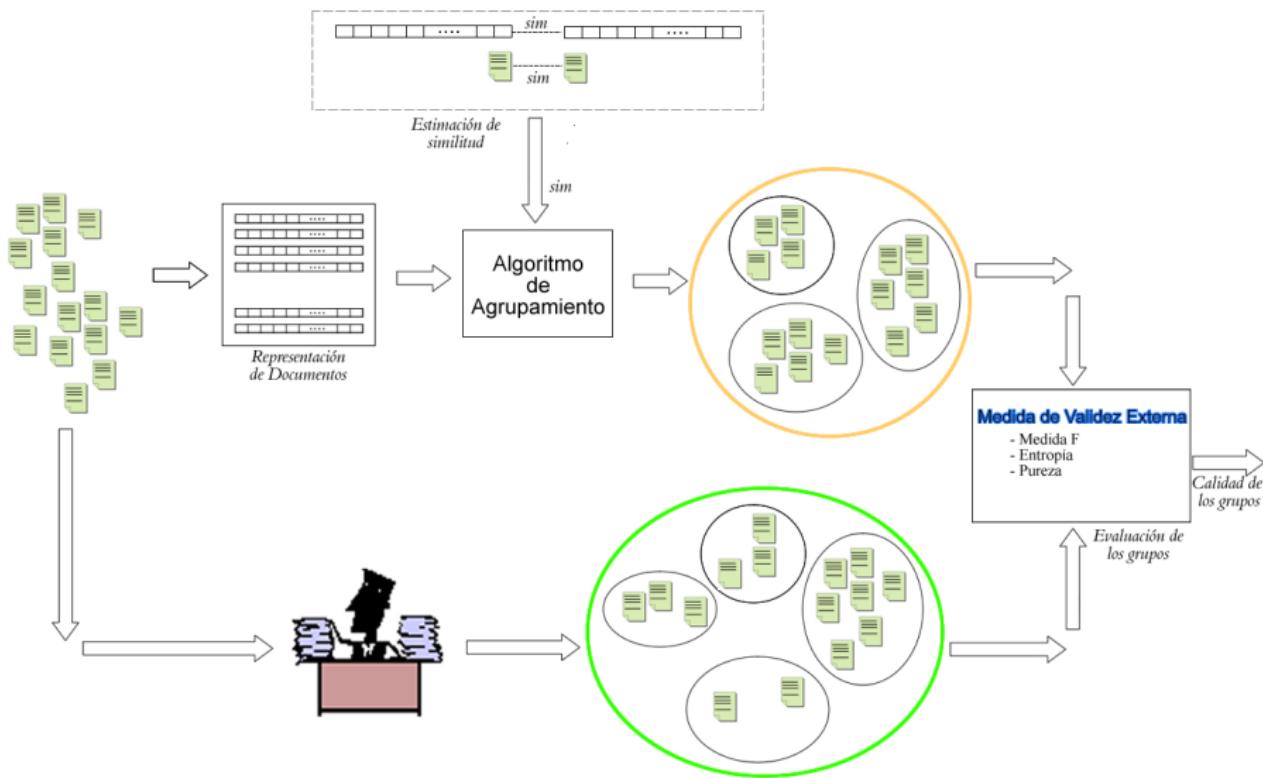
Agrupamiento de documentos



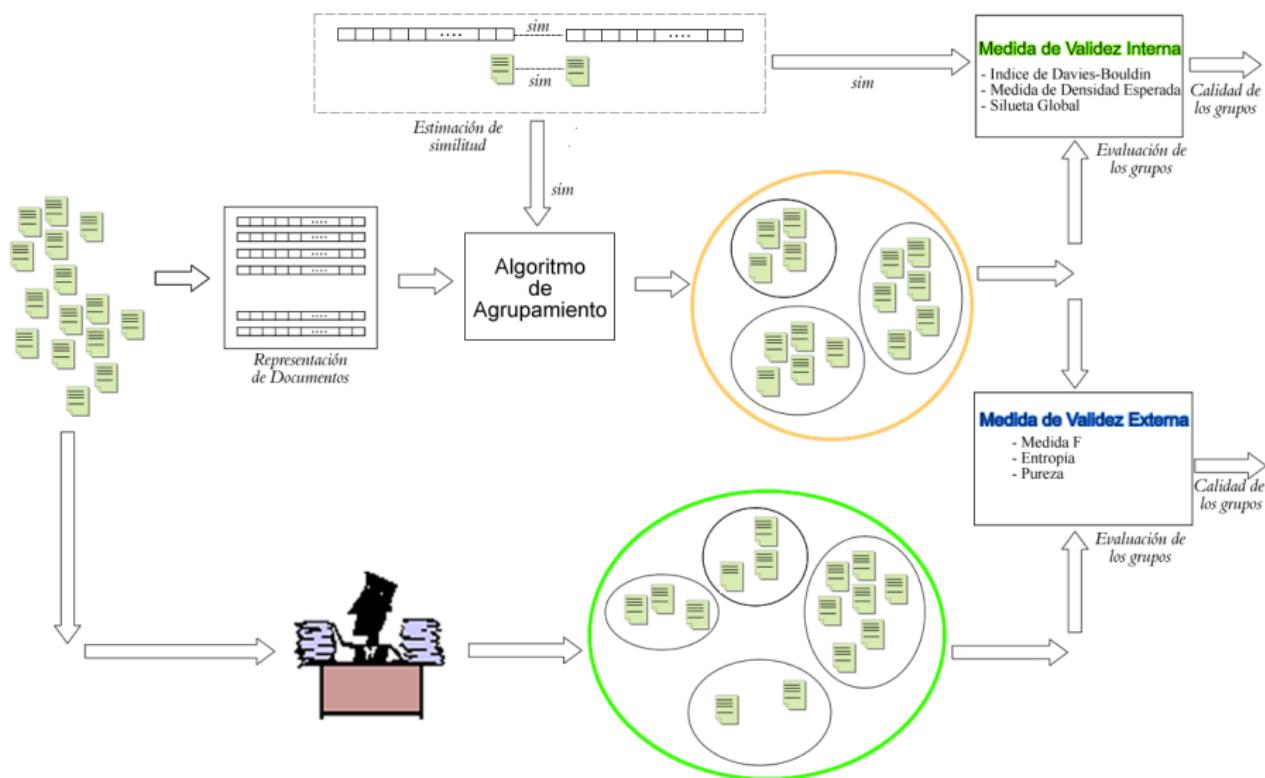
Agrupamiento de documentos



Agrupamiento de documentos



Agrupamiento de documentos



Extracción de Información (EI)

- Tarea que convierte texto **no (o semi) estructurado** en **representaciones estructuradas** (ej., **estructuras tabulares** para bases de datos).
- Las estructuras obtenidas contienen **información útil** y **etiquetada** referida a **nombres/entidades, relaciones** y **eventos** identificados en el texto original.
- Esta tarea, se difunde y gana interés creciente a partir de las conferencias “**Message Understanding Conferences (MUC)**” patrocinadas por el gobierno de los Estados Unidos

Ejemplo de EI

EI de actividad terrorista a partir de un cable de noticia.

Dado el siguiente texto:

19 March – A bomb went off this morning near a power tower in San Salvador leaving a large part of the city without energy, but no casualties have been reported. According to unofficial sources, the bomb – allegedly detonated by urban guerrilla commandos – blew up a power tower in the northwestern part of San Salvador at 0650 (1250 GMT).

Ejemplo de EI (II)

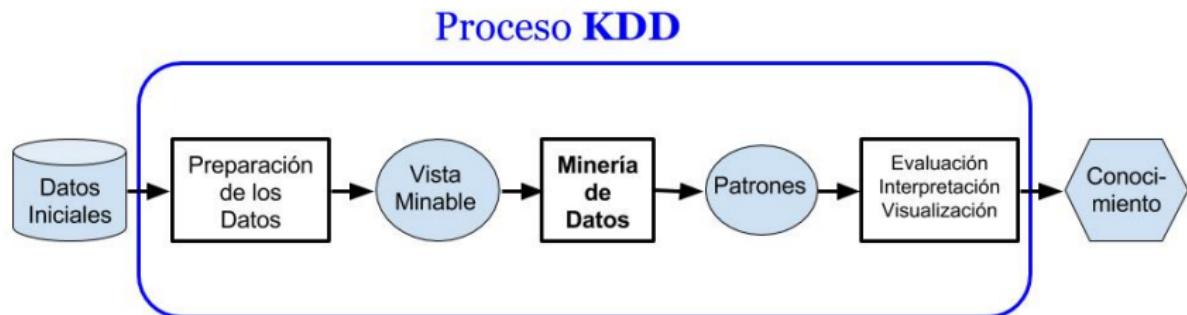
... se genera (“rellena”) el siguiente “template”:

Incident type	bombing
Date	March 19
Location	El Salvador: San Salvador (city)
Perpetrator	urban guerilla commandos
Physical target	power tower
Human target	-
Effect on physical target	destroyed
Effect on human target	no injury or death
Instrument	bomb

KDD a partir de textos

Sirve de guía para la organización de los contenidos de este curso.

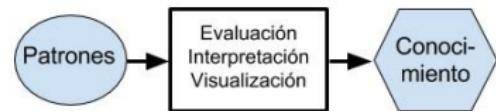
Fases del Proceso KDD



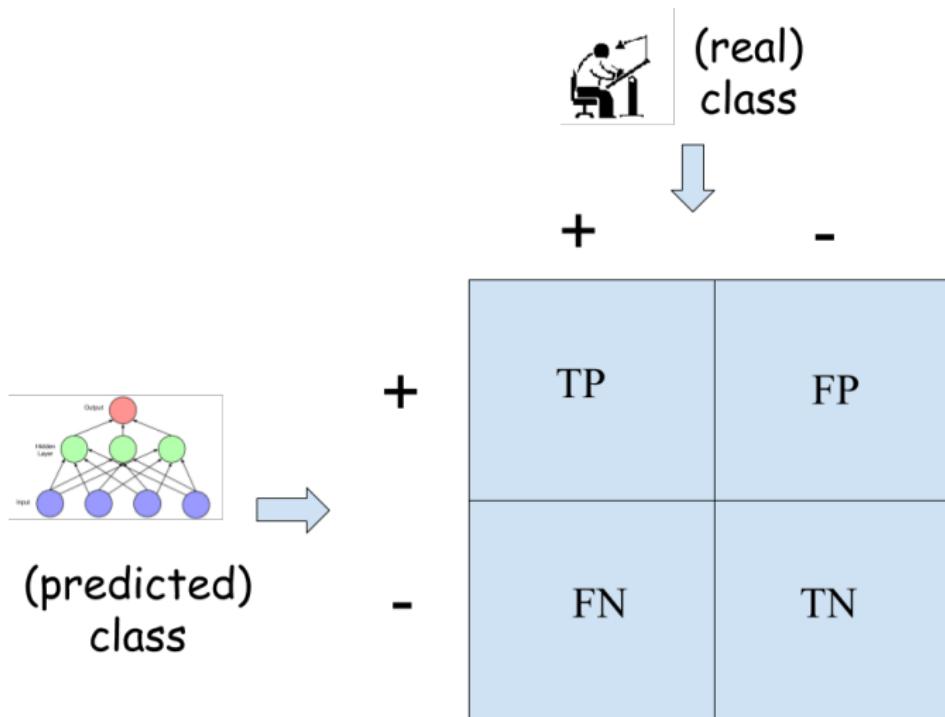
Fase de evaluación, interpretación y visualización

Sirve de guía para la organización de los contenidos de este curso.

Fases del Proceso KDD

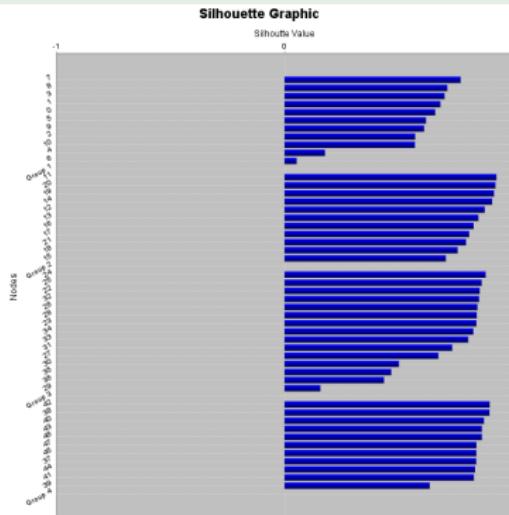


Evaluación (supervisada) clásica

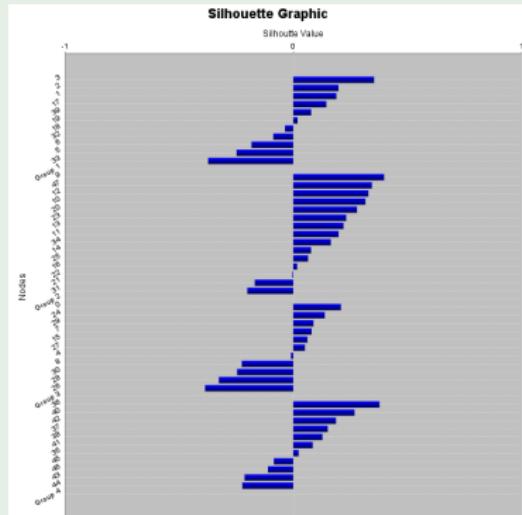


Evaluando agrupamientos

Agrupamiento bueno

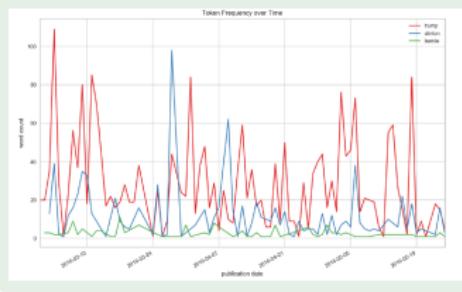


Agrupamiento malo

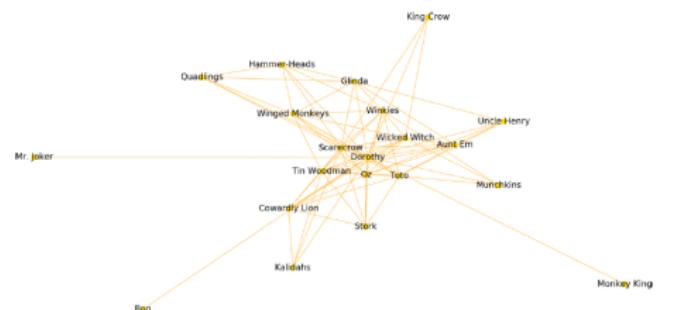


Visualizando información de textos

Frecuencia de palabras en el tiempo

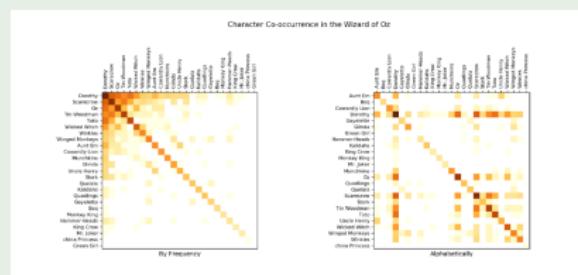


Grafo social de palabras

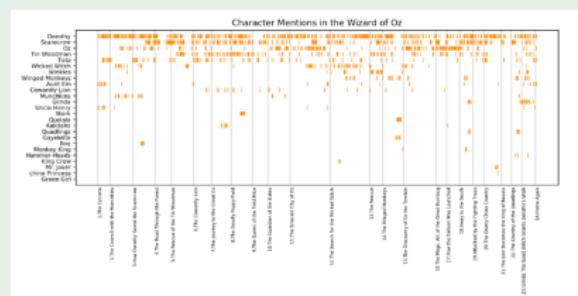


Visualizando información de textos

Matrices de co-ocurrencia

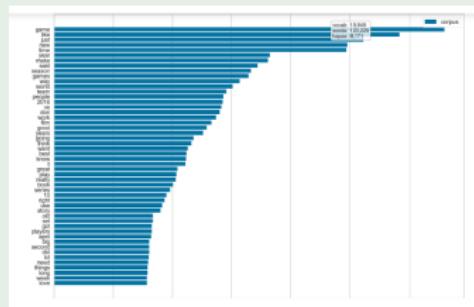


Gráficas de dispersión

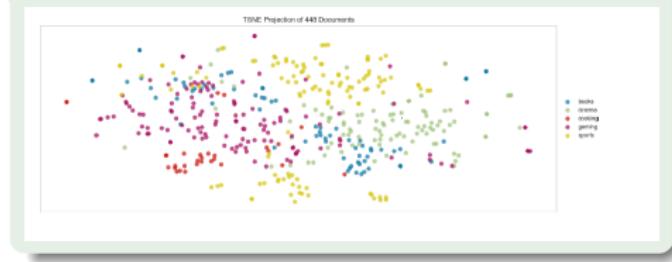


Visualizando información de textos

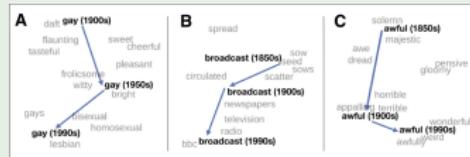
Frecuencias



Visualización con t-SNE



Evolución en el tiempo



Cercanía de embeddings



Niveles del Lenguaje Natural

La lingüística construye **modelos del lenguaje**.

Sin embargo, no construye modelos del lenguaje completo, sino que trata de hacerlo considerando sus partes más simples:

- ① Fonética/Fonología
- ② Morfología
- ③ Sintaxis
- ④ Semántica
- ⑤ Pragmática
- ⑥ Discurso

Niveles del Lenguaje Natural

Reflejan distintos **niveles de conocimiento** del lenguaje, necesarios en comportamientos del lenguaje complejo.

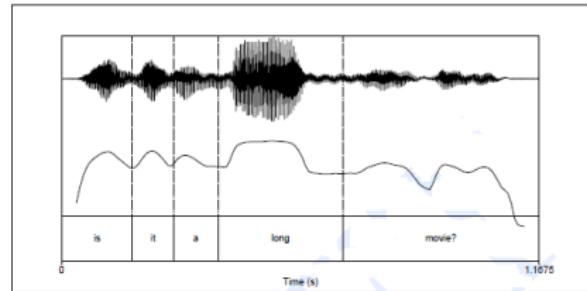
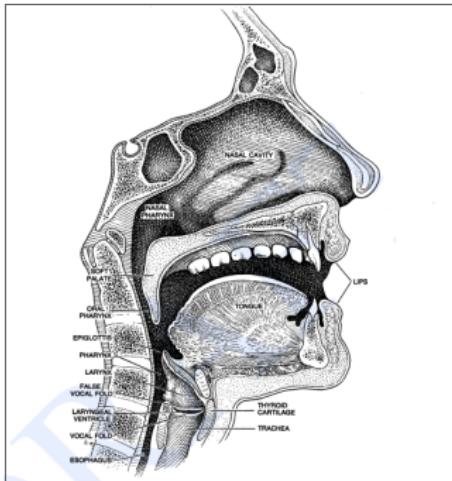
- ① **Fonética/Fonología:** conocimiento sobre los **sonidos** lingüísticos
- ② **Morfología:** conocimiento de las **componentes** significativas de las **palabras**
- ③ **Sintaxis:** conocimiento de las **relaciones estructurales** entre **palabras**
- ④ **Semántica:** conocimiento del **significado**
- ⑤ **Pragmática:** conocimiento de la relación entre el **significado** y los **objetivos** e **intenciones** del hablante
- ⑥ **Discurso:** conocimiento sobre las **unidades lingüísticas** más largas que una **oración simple**

Niveles del Lenguaje Natural

Fonética - fonología

Parte de la lingüística que se dedica a la exploración de las características del **sonido**

Sus métodos son en su mayoría **físicos**



Niveles del Lenguaje Natural

Morfología

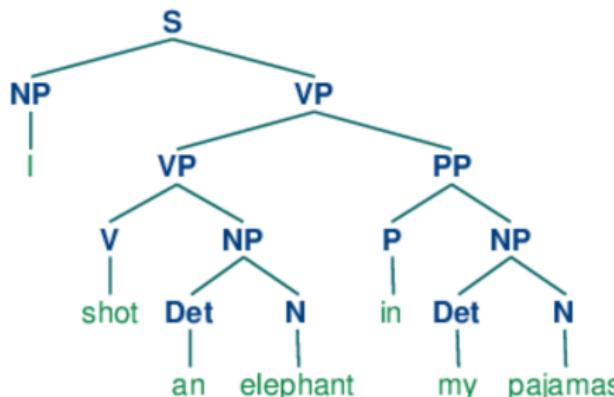
Se dedica a la **estructura interna** de las palabras (sufijos, prefijos, raíces, flexiones) y el sistema de **categorías gramaticales** de los idiomas (género, número, etc.).

- Se relaciona con funciones básicas del pre-procesamiento de textos como la **lematización** y el **etiquetado POS**
- La salida de un módulo morfológico es usualmente una tabla con los **lemas** y las **propiedades** de cada palabra del texto:
fuera ⇔ **SER**, subjuntivo, tercera persona, singular

Niveles del Lenguaje Natural

Sintaxis

- Se dedica a analizar las **relaciones** entre las **palabras** dentro de la **frase**.
- Un módulo **sintáctico** transforma la representación morfológica y genera una secuencia de **estructuras de oraciones** (**árboles sintácticos**)



Niveles del Lenguaje Natural

Semántica

- El propósito de la semántica es “entender” la frase.
- Esto involucra
 - Saber el **sentido** de todas las palabras
 - **Interpretar** las relaciones sintácticas.

Salidas usuales de un módulo semántico:

- **conceptos / sentidos** de las palabras
- **redes semánticas / grafos conceptuales** (representan todos los conceptos y las relaciones entre ellos)

Niveles del Lenguaje Natural

Pragmática

- La pragmática analiza las **intenciones** del autor del texto o del hablante (ejemplo de **pasar la sal**)
- También analiza las **clases de oraciones** que constituyen **acciones** en sí mismas (con **pre-condiciones** y **efectos**)

Discurso

- Analiza la relación entre **varias oraciones** a nivel del discurso completo.
- Problema importante: resolución de **coreferencias** (**anáforas**).
Ejemplo: "Juan llegó a la casa de Ana. Nunca pensó que ella ya había partido" (ella = Ana)

Niveles del Lenguaje Natural

Ejemplo: interacción de los niveles en **sistemas de diálogo**

