Introducción a las Técnicas Minería de Textos

Maria-Amparo Vila
vila@decsai.ugr.es

Grupo de Investigación en Bases de Datos y Sistemas de Información Inteligentes https://idbis.ugr.es/
Departamento de Ciencias de la Computación e Inteligencia Artificial
Universidad de Granada

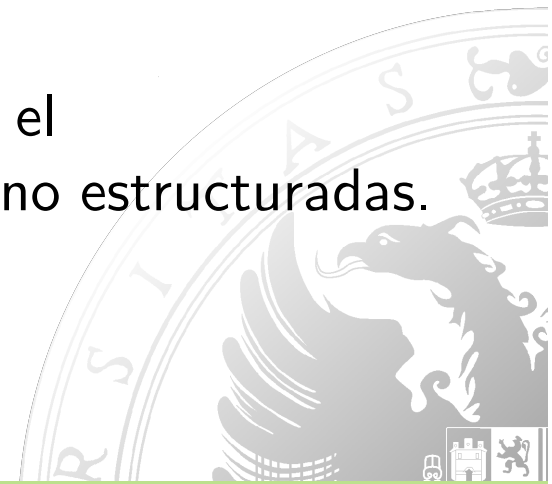
Esquema de la presentación

El problema de Minería de Textos

El propio concepto de "Minería de Textos" (TM) es algo que todavía se encuentra en discusión.
Se podría definir como:

> *"Proceso de extracción de conocimiento o patrones, previamente desconocidos, no triviales e interesantes (potencialmente útiles) y comprensibles por los usuarios a partir de documentos de texto no estructurados.".*

Text Mining es una extensión de Data Mining donde el descubrimiento se realiza a partir de Bases de Datos no estructuradas.



El problema de Minería de Textos

No se debe confundir Text Mining con Recuperación de Información a partir de bases de datos textuales.
La recuperación de información busca "documentos" de acuerdo con unos requerimientos. En TM buscamos:

- Conocimiento desconocido
- Comprensible por los usuarios
- No trivial
- Interesante

El problema de Minería de Textos

Se considera que la TM es el último paso en el procesamiento de textos:

El problema de Minería de Textos

Se podría pensar que se pueden aplicar directamente las técnicas clásicas de DM a la información textual. Nada más lejos de la realidad . La DM trabaja con Bases de datos con esquema conocido. Cada documento de texto es una colección ordenada de palabras y signos de separación con significado asociado cuya situación en el texto esta determinada por restricciones de tipo sintáctico y semántico. Existen textos semiestructurados tales como los documentos escritos en XML.
En TM, los datos son:

- Inherentemente desestructurados
  - Estructura implícita
  - Mucha mayor riqueza que en los casos estructurados
- Ambiguos
- Multilinguales

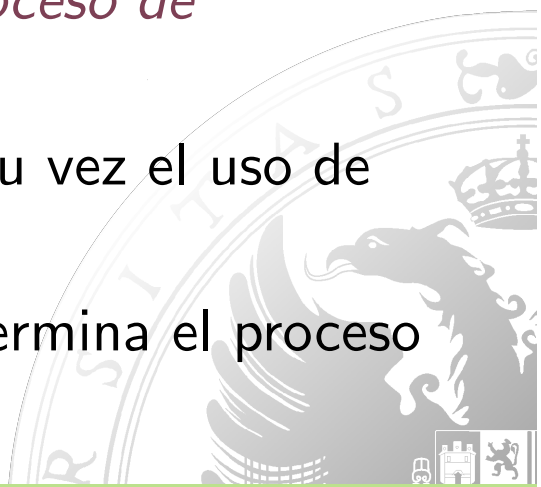
El problema de Minería de Textos

⋆ Esta ausencia de estructura es el mayor problema de la TM e implica la necesidad de preprocesar los textos, de pasarlos a una **forma intermedia**

- Bolsas (bags) de términos
- Estructuras matriciales (datasets)
- Grafos conceptuales o redes semánticas.
- Estructuras de tipo "ontología"

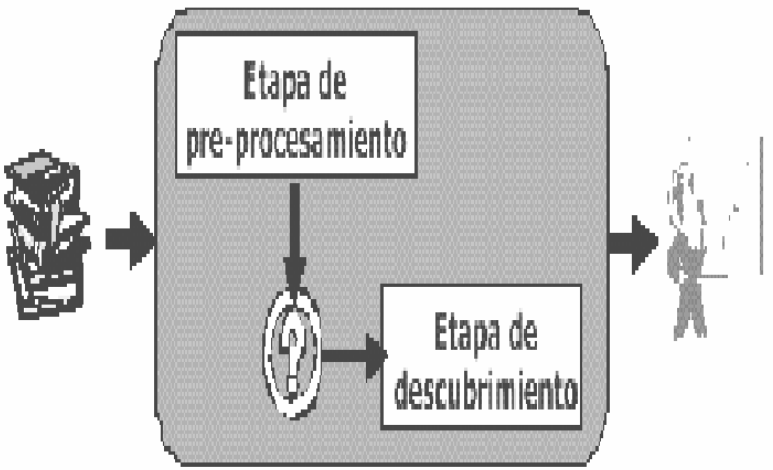
La fase de preparación de datos, inherente a todo proceso de extracción de conocimiento, es crucial.

- La obtención de la "forma intermedia" implica a su vez el uso de técnicas de extracción de conocimiento,
- En muchos casos, no esta claro entonces, donde termina el proceso de preparación de datos y empieza el de minería.



El problema de Minería de Textos

Un esquema sencillo de este proceso sería:



El problema de Minería de Textos

⋆ Algunos autores se limitan a definir la TM como un campo interdisciplinario que incluye elementos de:

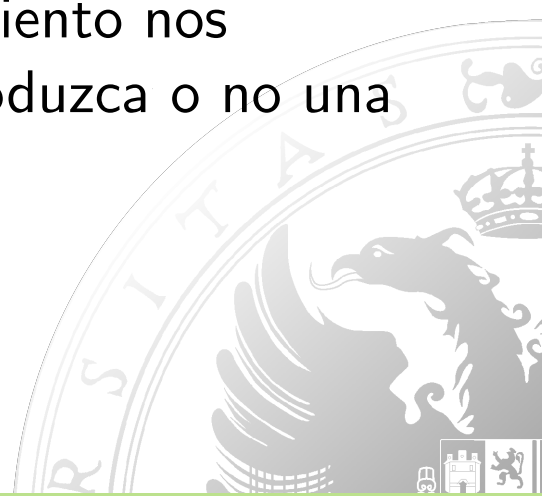
- Recuperación de información
- Extracción de información mediante lingüistica computacional
- Agrupamiento (clustering)
- Categorización,
- Otras técnicas de DM

El problema de Minería de Textos

⋆ Todos autores están de acuerdo en que el proceso de TM incluye las siguientes fases:

1. Preprocesamiento
2. Minería (propiamente dicha)
3. Visualización

En función de complejidad de la fase de preprocesamiento nos encontramos con la posibilidad de que esta etapa produzca o no una forma intermedia compleja.



Preprocesamiento en Minería de textos

Preprocesamiento Sintáctico

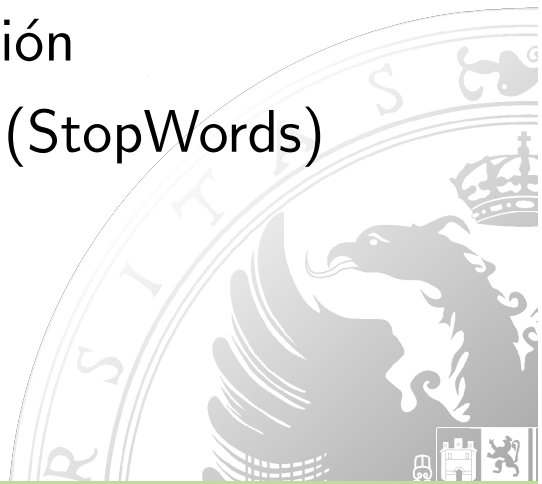
Idea básica

Se trata de procesar texto libre de manera que la salida pueda ser tratada de forma automatizada. Hay que pasar de datos no estructurados a una estructura de datos.

En principio Una bolsa de términos, posiblemente anotados

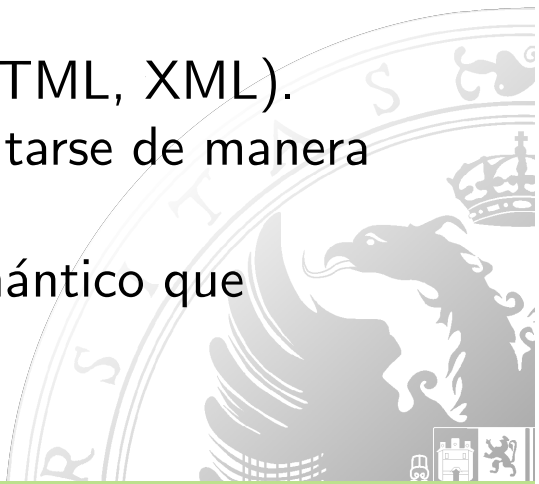
Etapas en el procesamiento sintáctico

1. Tokenización
2. Reconocimiento/Eliminación de signos de puntuación
3. Reconocimiento/Eliminación de "palabras vacías" (StopWords)
4. Reconocimiento de palabras múltiples. (n-gramas)
5. Reconocimiento de tipos sintácticos (POS)
6. Lematización (Steaming)



Tokenización

- Se trata de partir el texto en "tokens", cadenas de caracteres que representan palabras.
- Varios "tokens" representan a un mismo término o palabra.
- Al final tendremos bolsas de palabras con su ocurrencia en cada término
- Problemas habituales:
  - Preparar el documento para extraer el texto (pdf, HTML, XML). Distintas zonas del documento pueden tener que tratarse de manera diferente.
  - Palabras "compuestas", exige preprocesamiento semántico que veremos posteriormente



Reconocimiento/eliminación de signos de puntuación

Problemas:

- Hay que detectar cuando un signo es verdaderamente de puntuación
  - ()¡¿¡?¿" son delimitadores y pueden ser tokens.
  - - ' , ; y . pueden formar parte de un término o ser delimitadores.
  - Probablemente habrá que hacer un análisis en profundidad, si existen muchos signos.
- Es muy dependiente del idioma
- Al final del proceso se eliminan del conjunto de términos del documento.

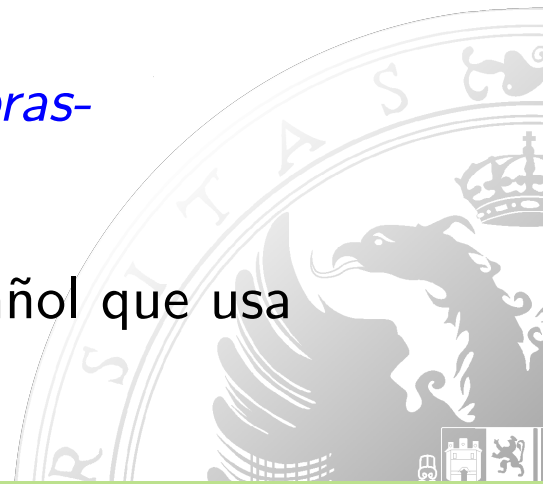
Reconocimiento/Eliminación de "palabras vacías" (StopWords)

- Son palabras que no proporcionan información desde un punto de vista no lingüísitico
- Tienen un papel esencialmente funcional
- Se eliminan del texto según una lista dada
- Es muy dependiente del lenguaje

En

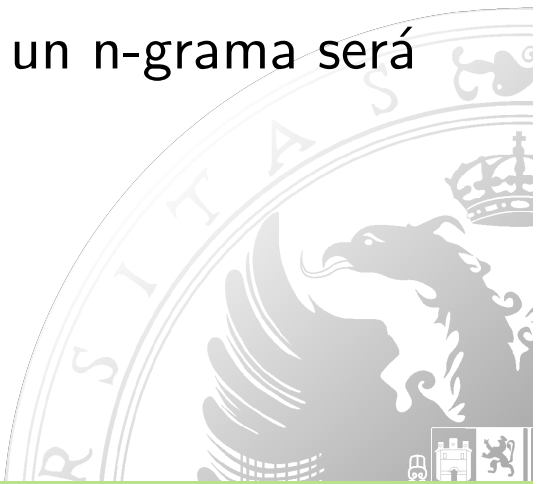
http://www.navigla.es/posicionamiento-seo/palabras-stopwords-seo-espanol/

se puede encontrar la lista de palabras vacías en español que usa google en su motor de búsqueda



Reconocimiento de palabras múltiples.

- Existen grupos de palabras que tiene significado por si mismos: Data base, Sistema Operativo etc. estos grupos hay que tratarlos como términos. Pueden considerarse como un caso particular de n-gramas (n-grams)
- Su detección automática se basa en la idea de que un n-grama será más frecuente que n palabras juntas cualquiera.



Preprocesamiento en Minería de textos

Preprocesamiento Sintáctico

Reconocimiento de palabras múltiples. Para detectar n-gramas un posible algoritmo:

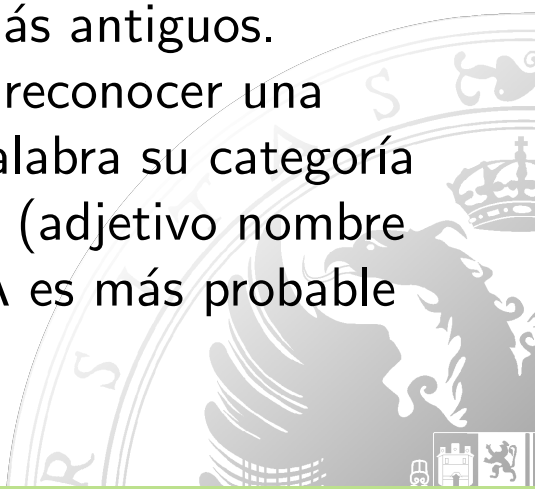
1. Fijar un valor de n (2,3,..) máximo de palabras para ser consideradas conjuntamente. Fijar un umbral de medida de aparición de n-gramas: Minmes
2. Para k=2..n
   - 2.1 Se analizan las secuencias de k-palabras consecutivas en el texto T_k.
   - 2.2 Se seleccionan aquellas T_k tales que:

$$AM(T_k) = \frac{k(log_{10} frec(T_k))(frec(T_k)}{\sum_{word_i \in T_k} frec(word_i)} \geq minmes$$

   - 2.3 Si existe algún $T_{k-1}, k > 2$ que esté incluido en T_k eliminar $T(k-1)$ del conjunto de n-gramas seleccionado

Reconocimiento/Eliminación de tipos sintácticos (Part of Speech, POS)

- Para ciertas aplicaciones es necesario reconocer las funciones gramaticales de los términos: nombre, nombre propio, verbos en distintos tiempos etc.
- Existen algoritmos que permiten etiquetar los términos en este sentido.
  - Algoritmos basados en reglas lingüísticas. Son los más antiguos.
  - Algoritmos de aprendizaje automático. Se basan en reconocer una categoría para un término y asignar a la siguiente palabra su categoría más probable según el lenguaje. P.E. en ingles ANN (adjetivo nombre nombre) es una secuencia probable, en español NNA es más probable
- Obviamente dependen del idioma



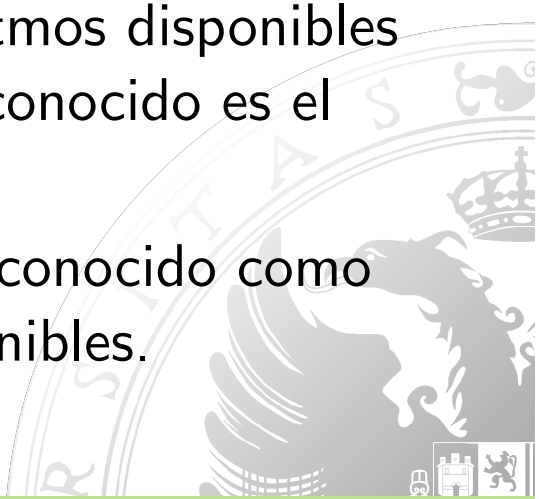
Reconocimiento/Eliminación de tipos sintácticos (Part of Speech, POS)

Algunas categorias de POS en inglés

Tag	Description
CC	Coordinating conjunction
CD	Cardinal number
DT	Determiner
EX	Existential there
FW	Foreign word
IN	Preposition or subordinating conjunction
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
LS	List item marker
MD	Modal
NN	Noun, singular or mass
NNS	Noun, plural
POS	Possessive ending
UH	Interjection
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund or present participle
VBN	Verb, past participle
VBP	Verb, non-3rd person singular present
WDT	Wh-determiner

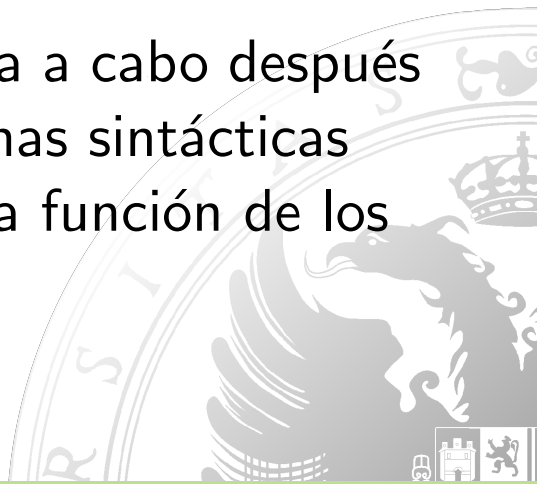
Lematizacion (Steaming)

- Las diferentes formas una misma palabra suelen ser problemáticas para el análisis de textos, ya que tienen diferente ortografía y significado similar (por ejemplo, aprende, aprender, aprendizaje)
- Entendemos por *Lematización (Steaming o Lematization)* el proceso transformar una palabra en su "raiz standarizada"
- Para Inglés no es un gran problema, existen algoritmos disponibles públicamente que dan buenos resultados. El más conocido es el algoritmo de Porter
- Para español no hay un algoritmo generalmente reconocido como el de Porter pero existen diversos algoritmos disponibles.



Lematizacion (Steaming)

- No siempre es útil llevar a cabo la lematización.
  - En general para los casos en que se trabaje con frecuencia de términos es interesante relizarlo ya que resume varios términos en un sólo y aumenta frecuencia de este.
  - Pero puede dar problemas si se hace un preprocesamiento semántico posterior
- En cualquier caso la lematización habrá que llevarla a cabo después del etiquetado y eliminación en su caso de las formas sintácticas (POS) no deseadas, pues la lematización cambia la función de los terminos.



Lematizacion (Steaming)

Reglas de lematización en inglés

- ATIONAL -> ATE relational -> relate
- TIONAL -> TION conditional -> condition
- ENCI -> ENCE valenci -> valence
- ANCI -> ANCE hesitanci -> hesitance
- IZER -> IZE digitizer -> digitize
- ABLI -> ABLE conformabli -> conformable
- ALLI -> AL radicalli -> radical
- ENTLI -> ENT differentli -> different
- ELI -> E vileli -> vile
- OUSLI -> OUS analogousli -> analogous

Preprocesamiento en Minería de textos

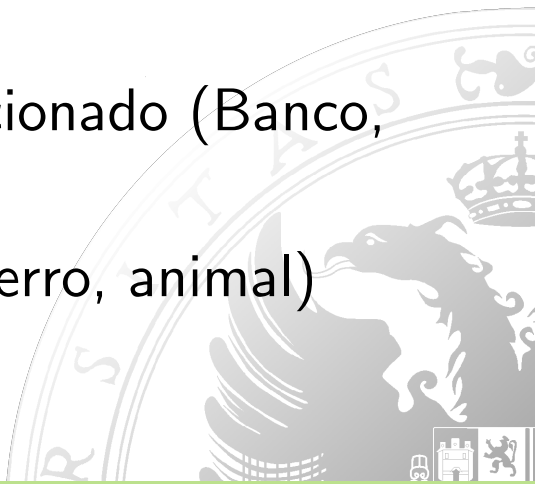
Preprocesamiento Semántico

Idea básica

Una vez limpios y etiquetados los términos se utilizan relaciones de tipo semántico para reducirlos nuevamente.

Posibles relaciones entre los términos

- *Sinonimia*: distinta forma, igual significado (clase, lección)
- *Homonimia*: misma forma distinto significado (banco institucion financiera, sitio de sentarse)
- *Polisemia*: misma forma ,distintos significado relacionado (Banco, Banco de sangre)
- *Hiponimia* Una palabra es una subclase de otra (perro, animal)

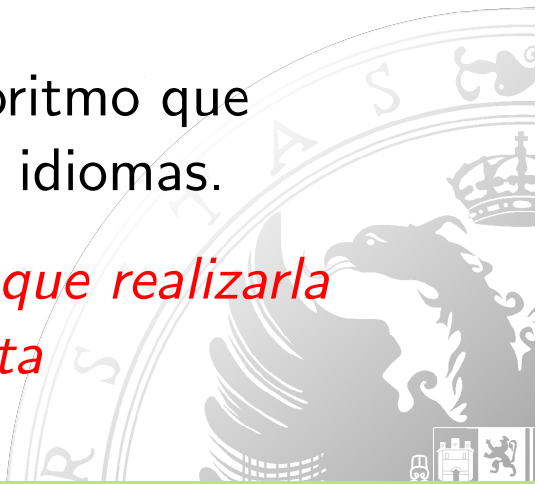


Desambiguación

Proceso mediante el cual se asigna a varios términos uno sólo que tiene el mosmo significado que todos ellos

- Existen herramientas que ayudan a trabajar con sinónimos e hipónimos, llegando a asignar cada conjunto de terminos a una clase semántica (desambiguación).
- La más famosa es Wordnet; pero no existe un algoritmo que resuelva totalmente el problema y menos en varios idiomas.

La desambiguación es un problemas abierto y hay que realizarla "artesanalmente" para cada problema, si se necesita

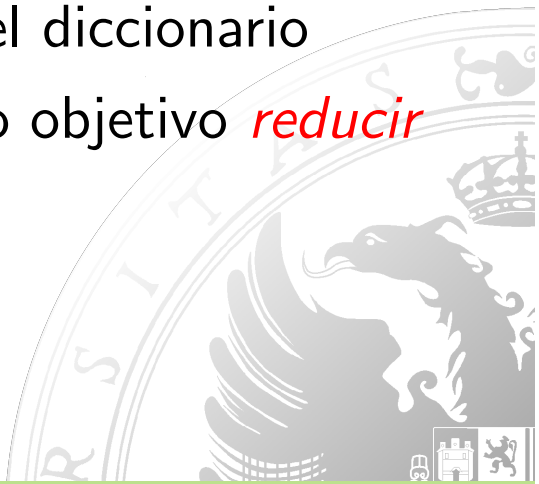


Reducción de variables

Ideas básicas

Problemas

- Dada una gran colección de documentos cada uno de ellos está¡ caracterizado por sus términos.
- El conjunto de términos de una colección se denomina **diccionario** y puede tener varios miles de elementos.
- La representación de cada documento se basa en el diccionario
- Las técnicas de reducción de términos tienen como objetivo *reducir el número de términos del diccionario*



Reducción de variables

Ideas básicas

Técnicas directas

Se usan directamente en la fase de preprocesamiento. Reducen el diccionario:

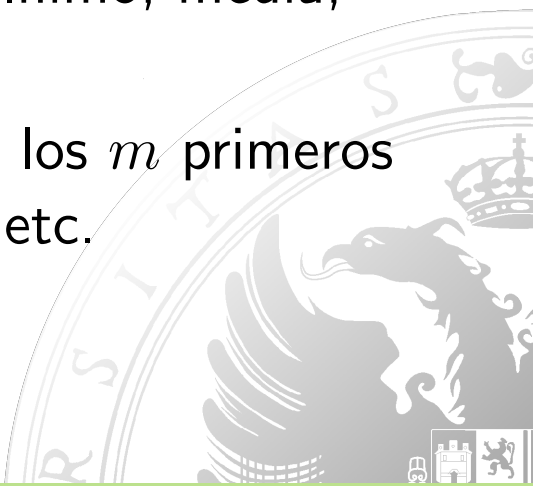
1. La eliminación de palabras vacías
2. El reconocimiento de palabras múltiples
3. La eliminación de tipos sintácticos no significativos
4. La lematización
5. La desambiguación

Reducción de variables

Técnicas basadas en medidas

Proceso básico

- Para cada término k y cada documento i se establece una "medida de importancia" w_{ik} del término k en el documento i.
- Para medir la importancia de un término en una colección de documentos se establece $w_k = Agg_{i=1}^{n} w_{ik}$ donde Agg es una medida de agregación que puede ser el máximo, mínimo, media, suma etc.
- Se ordenan los términos según w_k y se seleccionan los m primeros o un porcentaje sobre el total (p.e. el 75% mejor) etc.



Medidas basadas en frecuencias

- Sea n_{ik}= número de veces que aparece el término k en el documento i, y n_i número de términos asociados a i
- La frecuencia de cada término k en i viene dada por $f_{ik} = \frac{n_{ik}}{n_i}$
- Si trabajamos con medidas de frecuencias tomamos $w_{ik} = f_{ik}$. w_k puede ser: $w_k = t_k = \sum_{i=1}^{n} f_{ik}$ frecuencia total o bien $w_k = \sum_{i=1}^{n} f_{ik}/n$ frecuencia media.

Reducción de variables

Técnicas basadas en medidas

Medidas de discriminación

Problema

¿Son los términos más frecuentes los que mejor representan una colección de documentos?. No representarán mejor ciertos términos que aparezcan en algunos documentos y no en otros, que discriminen una parte de los documentos?

Medida tf*idf Si d_k es el número de documentos en el que aparece k, cuanto mayor sea menos discrimina k. Se define entonces:

$$idf_k = log_2(N/d_k) + 1 \text{ y } w_{ik} = tf * idf_{ik} = f_{ik} idf_k$$

Reducción de variables

Técnicas basadas en medidas

Medidas de discriminación

Medida del ruido Para cada término se define:

$$n_k = \sum_i (f_{ik}/t_k) Log_2(f_{ik}/t_k) \text{ y } s_k = log_2 t_k - n_k$$

$$\text{entonces } w_{ik} = f_{ik} s_k$$

Medidas basadas en similitud de documentos Para una colección de documentos se define la **similitud media entre ellos** σ, para cada término k se define σ_k como la *similitud media suprimiendo el término k de los documentos*. $\delta_k = \sigma_k - \sigma$ mide el poder discriminador de k.

$$w_{ik} = f_{ik} \delta_k$$

Reducción de variables

Técnicas basadas en selección por componentes principales

Modelo inicial :*El modelo vectorial documentos/términos*

Todo documento se representa como un vector: $d = (w_1, ..w_M)$ donde cada w_i representa el peso que tiene el término i en el documento d.

Problema

Aunque se haya realizado un *proceso de limpieza* de "palabras vacias", eliminación de sinónimos etc.., y un *proceso de reducción por frecuencia* podemos tener cientos de variables, es decir M puede ser muy grande. Mucho pesos pueden ser cero en un determinado documento, es decir, los vectores son muy grandes y "vacios"

Hay que hacer un proceso de reducción de dimensiones

El proceso de reducción de dimensiones transforma un vector $(w_1, ..w_M)$ donde d es muy grande en otro $(v1, ..v_m)$ donde m puede fijarse.

Reducción de variables

Técnicas basadas en selección por componentes principales

El problema

Dada una variable n dimensional $\bar{x} = (x_1, ..x_n)$ y m valores de la misma $x_{ij}, i \in \{1, , m\}, j \in \{1, .., n\}$, encontrar transformaciones lineales normalizadas (SLC) que "resuman" lo mejor posible los datos,capturando la mayor varianza de los mismos.

Idea Intuitiva

Si se consideran los items como una nube de puntos en R^n, todos ellos se pueden encerrar en un elipsoide, de centro la media, cuya matriz es la matriz de covarianzas.
Los ejes del elipsoide son un sistema de coordenadas ortogonal , si realizamos un cambio a este sistema de coordenadas, los puntos de dispersan a lo largo de los ejes

Reducción de variables

Técnicas basadas en selección por componentes principales

El modelo matemático

Sea $\bar{\mu}$ la media de $\bar{x}$ y Σ su matriz de covarianza, se trata de encontrar una transformación lineal $\bar{y} = \Gamma'(\bar{x} - \bar{\mu})$ tal que los nuevos ejes de coordenadas sean los ejes del elipsoide. Se prueba que Γ es una matriz tal que:

$$\Gamma'\Sigma\Gamma = \Lambda$$

donde,

$$\Lambda = diag(\bar{\lambda}) = \begin{pmatrix} \lambda_1 & & & & \\ & \ddots & & & \\ & & \ddots & & \\ & & & \ddots & \\ & & & & \lambda_n \end{pmatrix}$$

Reducción de variables

Técnicas basadas en selección por componentes principales

Los valores del vector $\bar{\lambda}$ verifican $\lambda_1 \geq, .. \geq \lambda_n$ y son los "autovalores" de la matriz de covarianza y asociado a cada uno de ellos λ_j existe un "autovector" $\bar{\gamma_j}$ que es la j-esima columna de la matriz Γ y verificándose:

$$\forall j \in \{1, , n\}\ y_j = \bar{\gamma_j}'(\bar{x} - \bar{\mu})$$

y_j se denomina j-esimo componente principal.

$$\forall k, j \in \{1, , n\}\ Cov(y_j, y_k) = 0\ Var(y_j) = \lambda_j$$

$$Var(y_1) \geq .. \geq Var(y_n)$$

Reducción de variables

Técnicas basadas en selección por componentes principales

Proporción de varianza explicada

$\star$ La proporción de varianza explicada por k factores es $(\lambda_1 + .. + \lambda_k)/\lambda_1 + .. + \lambda_n)$ y nos permite reducir la dimensionalidad del espacio. Es decir expresar el fenómeno con menos variables.

Cuantas componentes tomar?:

- Al menos el 90% de varianza explicada
- Todos los autovalores que sean mayores que la media de los mismos.
- Si se utiliza la matriz de correlación en lugar de la matriz de covarianzas autovalores mayores que 1.

La proporción de variación explicada de la variable j por la componente k esta dada por $\rho(x_j, y_k) = r_{ik}$, con estos valores se pueden identificar los componentes e identificar una semántica para ellos.

Reducción de variables

Técnicas basadas en selección por componentes principales

Nueva representación de documentos y términos

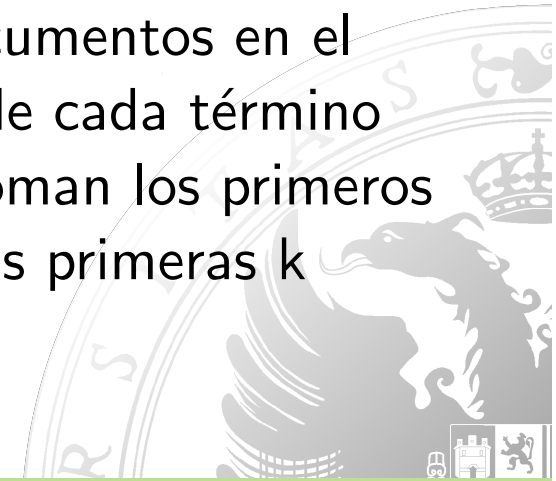
De la expresión:

$$\bar{y} = \Gamma'(\bar{x} - \bar{\mu})$$

podemos obtener:

$$\bar{x} = \Phi\Lambda\Delta'$$

De forma que, si Λ representa la matriz de los nuevos "factores", entonces Δ representa la representación de cada documentos en el nuevo espacio de los factores y Φ la representación de cada término en función de los factores. Lo mismo ocurres si se toman los primeros k elementos de Λ, las primeras k columnas de Φ y las primeras k columnas de Δ.



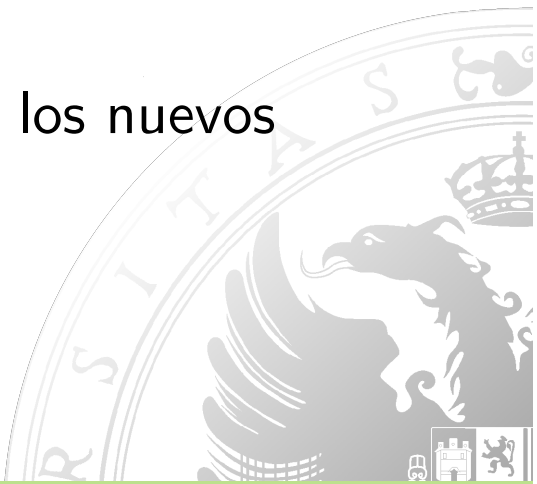
Reducción de variables

Técnicas basadas en selección por componentes principales

Análisis de semántica latente

En nuestro caso el modelo se identifica cómo:

- Los items son los documentos
- Los pesos de los términos en cada documentos son las variables
- Los nuevos factores son combinaciones de terminos relacionados entre si
- Las representaciones de documentos en función de los nuevos factores pueden utilizarse de formas muy diversas



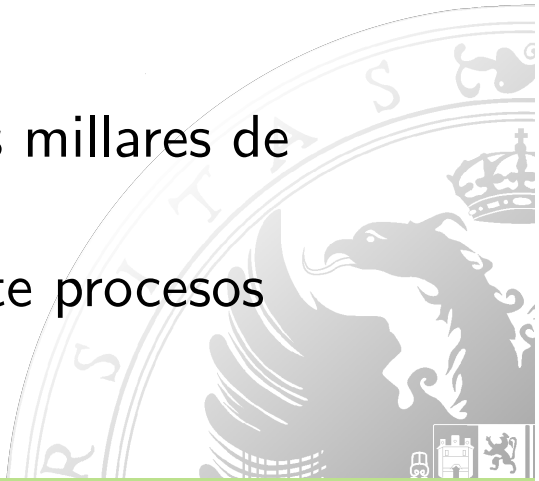
Reducción de variables

Técnicas basadas en selección por componentes principales

Análisis de semántica latente

- las representaciones gráficas segun factores pueden dar una idea de un primer agrupamiento de documentos
- Las representaciones de términos en función de los factores pueden dar una idea de la semántica de los mismos. Hasta el punto que algunos autores hablan de nuevos "conceptos" identificados mediante los factores.

Experiencias previas muestran que:
- Se pueden reducir a 200 factores, espacios de varios millares de términos conservando más de un 90% de la varianza
- La reducción de dimensiones no afecta sensiblemente procesos posteriones de TM como el agrupamiento



Agrupamiento y TM: algunos enfoques

Modelo clásico

- Datos de partida: *modelo vectorial o vectorial reducido*
- Distancia utilizada: *La medida del coseno*
  Ya que cada documento es un vector se calcula en coseno del ángulo que forman . Si $t_1 = (w_{11}...w_{1d})$ y $t_2 = (w_{21}...w_{2d})$ son dos vectores, entonces:

$$cos(t_1, t_2) = (t_1 \odot t_2)/|t_1||t2|$$

  donde $\odot$ representa el producto escalar y $|.|$ el módulo, es decir:

$$cos(t_1, t_2) = \frac{\sum_{j=1}^{d} w_{1j} w_{2j}}{\sqrt{\sum_{j=1}^{d} w_{1j}^2} \sqrt{\sum_{j=1}^{d} w_{2j}^2}} \quad (1)$$

- Si se utiliza un modelo de factores se puede usar la distancia euclídea sin problemas

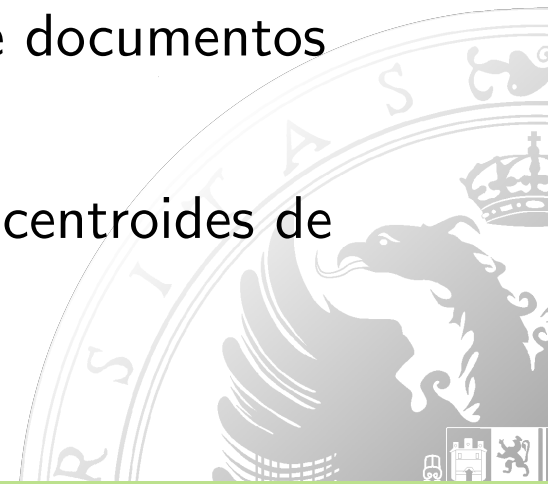
Agrupamiento y TM: algunos enfoques

⋆ Es muy habitual utilizar el **método de las k-medias** con la distancia del coseno. Se emplean algunas variantes que permiten la mejora del método:
-Selección de centroides mediante

- Un cluster jerárquico inicial en un conjunto pequeño de documentos.
- Análisis en componentes previos y posterior representación
- Otras técnicas de fraccionamiento del conjunto de documentos

- Utilización de mejoras de las k-medias:

- Método de las k-medias continuo: se calculan los centroides de forma continua, no al final de cada paso

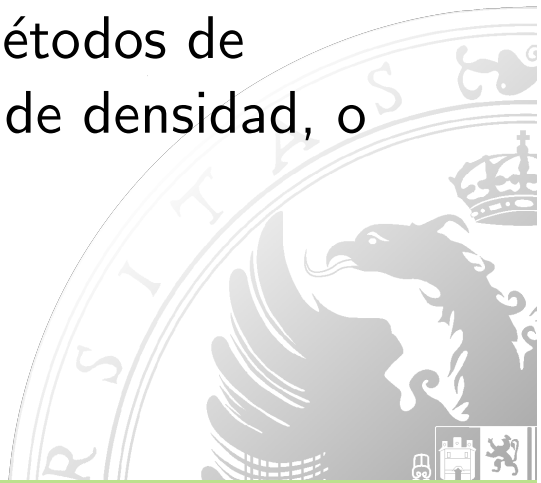


Agrupamiento y TM: algunos enfoques

- Utilización de mejoras de las k-medias:

- Método de las k-medias en dos fases, donde se realiza un refinamiento del proceso en cada etapa del algoritmo para evitar caer en "óptimo locales " de la función de coherencia.
- Método de las k-medias por bipartición.

$\star$ Existen enfoques que recomiendan el uso de técnicas más elaboradas de agrupamiento particional tales como métodos de medoides (CLARANS) o métodos basados en análsis de densidad, o métodos jerárquicos avanzados (BIRCH).



Agrupamiento y TM: descubrimiento de sucesos

Problema

Estamos rodeados por textos que "cuentan sucesos", es decir de "noticias". El descubrimiento de sucesos en el contexto de las noticias es la identificación de "relatos" que correspondan a sucesos nuevos o previamente no identificados. La detección de sucesos puede realizarse de dos maneras:

- De forma retrospectiva
- On line

Agrupamiento y TM: descubrimiento de sucesos

Nuevos sucesos ocurren de modo continuo y las noticias relativas a ellos habitualmente se presentan en forma de "reventones" o "explosiones" con las siguientes características:

1. Aparecen agrupadas en el tiempo,
2. Las "explosiones" asociadas a nuevos sucesos aparecen separadas por un cierto margen de tiempo,
3. Cada nuevo suceso suele ir acompañado de un cambio en el tipo de términos y el vocabulario empleado para describir las noticias asociadas,
4. Los sucesos suelen ocupar una ventana temporal de entre una y cuatro semanas de duración.

Estas cuatro observaciones parecen sugerir que cualquier técnica de agrupamiento debe dar buen resultado en la detección de nuevos sucesos, tanto en el caso retrospectivo como en el caso "on line".

Agrupamiento y TM: descubrimiento de sucesos

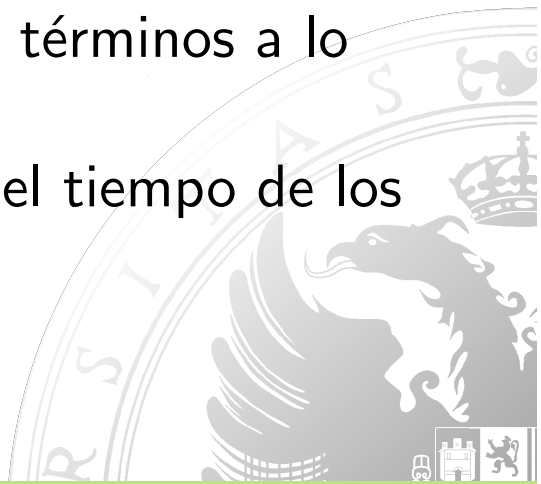
Modelo

- Cada noticia es un "objeto" caracterizado por un conjunto de atributos (términos)
- Cada suceso es un "prototipo" de un conjunto de noticias similares en función de sus términos.

En el caso de detección retrospectiva se buscan "explosiones" en un archivo histórico por medio de :

◇ Cambios bruscos en la distribución de términos a lo largo del tiempo

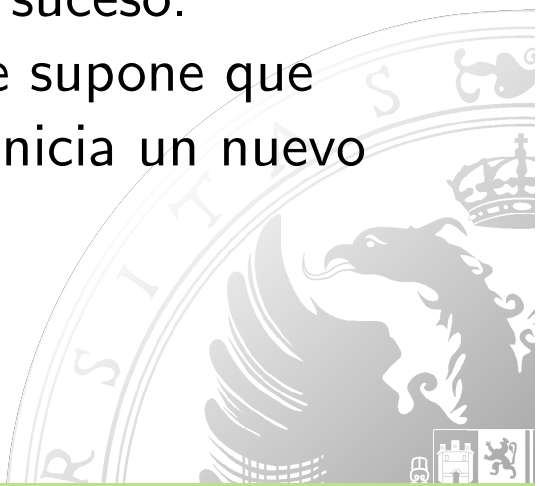
◇ Similaridades léxicas y proximidad en el tiempo de los textos de las noticias.



Agrupamiento y TM: descubrimiento de sucesos

En la detección "on line" la idea es procesar las noticias conforme van llegando. Los algoritmos que se han desarrollado para este problema son de respuesta "booleana":

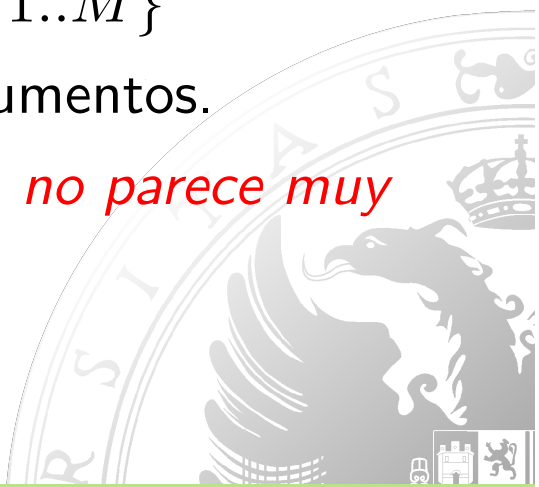
1. Cada noticia se compara, cuando aparece, con un "agrupamiento" ya detectado.
2. Si la noticia puede incorporarse a dicho agrupamiento se supone que corresponde al mismo suceso.
3. Si la noticia no puede incorporarse se supone que corresponde a un nuevo suceso y se inicia un nuevo agrupamiento.



Modelo general de clasificación

- Consideremos*El modelo vectorial documentos/términos*

  Todo documento se representa como un vector: $d = (w_1, ..w_N)$ donde cada w_i representa el peso que tiene el término i en el documento d.
- Supongamos que tenemos M documentos $\{d_i, i = 1..M\}$
- Un conjunto de H clases establecidas para los documentos.
- El problema de clasificar documentos (categorizar) *no parece muy diferente del problema de clasificación general*



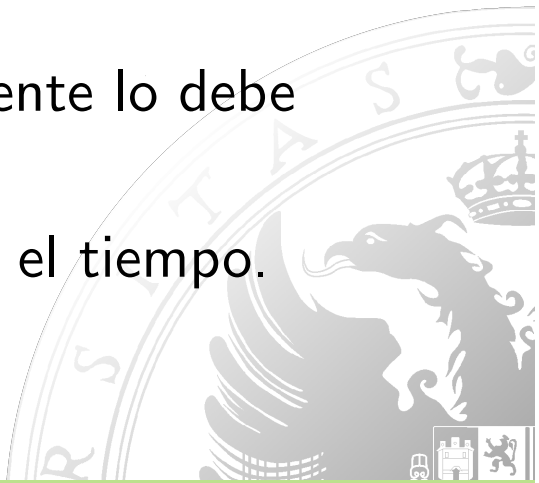
Problemas típicos de clasificación de textos

Categorización de documentos

Tenemos una colección de documentos clasificados por tópicos y queremos entrenar un categorizador de forma que nos permite clasificar nuevos documentos. Suelen tratarse de documentos "largos".

Problemas

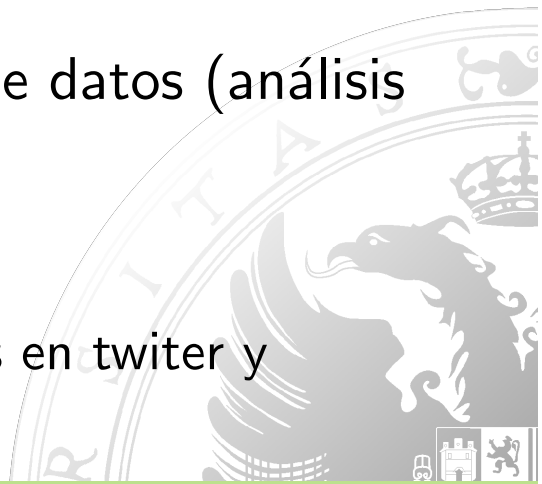
- Clasificar inicialmente los documentos. Habitualmente lo debe hacer un experto.
- Los tópicos de los documentos suelen cambiar con el tiempo.



Problemas típicos de clasificación de textos

Problemas con textos más cortos que incluyen un catgorizador de textos.

- Clasificador de noticias (Google news por ejemplo). Está relacionado con la detección de sucesos. En algunos casos se hace clustering previo y se trabaja con jerarquías
- Filtros de SPAM en correos electrónicos
- Análisis de sentimientos en redes sociales.
- Estudios de textos cortos categorizados en bases de datos (análisis descriptivos):
  - Descripciones médicas. Diagnósticos etc.
  - Encuestas de opinión con texto libre.
  - Minería Web de uso y contenido (análisis de tópicos en twiter y hábitos de uso etc..)

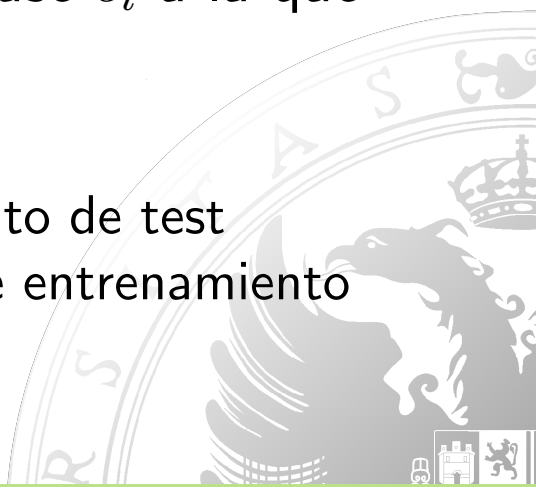


Modelo general de clasificación

- Tenemos el dataset:

items\variables	t_1	t_2	$\cdots\cdots$	t_N	C
d_1	w_{11}	w_{12}	$\cdots\cdots$	w_{1N}	c_1
$\vdots$	$\vdots$	$\vdots$	$\cdots\cdots$	$\vdots$	$\vdots$
d_M	w_{M1}	w_{M2}	$\cdots\cdots$	w_{MN}	c_M

- Se trata de predecir, con los valores de $\{w_{ij}\}$ la clase c_i a la que pertenece el documento.
- En principio el proceso de clasificación no varia:
  1. Tenemos un conjunto de entrenamiento y un conjunto de test
  2. Aplicamos un modelo de clasificación al conjunto de entrenamiento
  3. Aplicamos el modelo para predecir el conjunto test.
  4. Medimos la bondad del modelo



Clasificación de textos

Problemas específicos de clasificación de textos

Problema

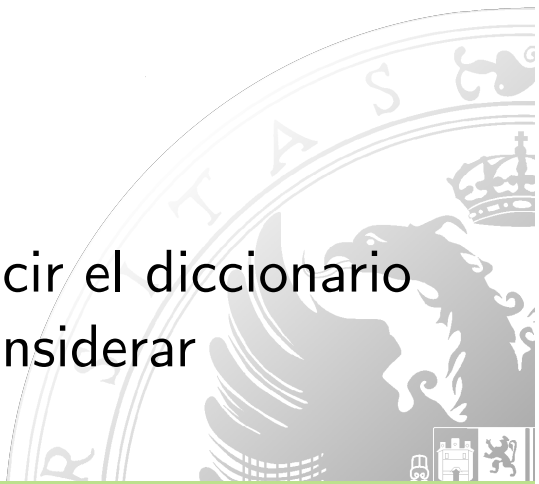
La representación vectorial de documentos:

1. En principio contiene muchas variables. Es imposible pensar, en principio, en técnicas discretas como árboles de decisión
2. La matriz documentos/términos es muy dispersa.

Es necesario **reducir drásticamente** *el espacio de términos*

Solución

1. Aplicar reducción en componentes principales.
2. Establecer procesos de clasificación binarios y reducir el diccionario a los términos más representativos de la clase a considerar



Clasificación de textos

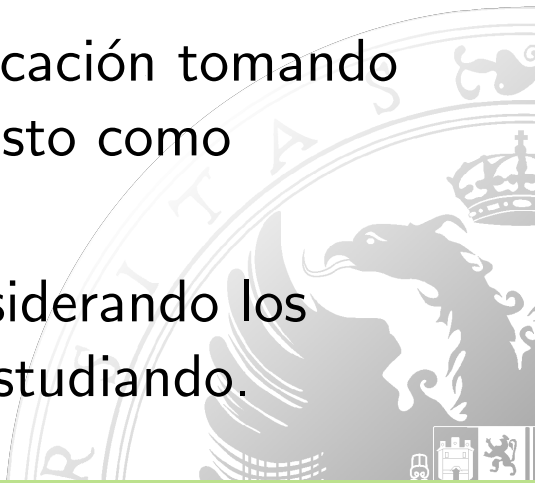
Problemas específicos de clasificación de textos

Problemas con la aplicación de componentes principales

- Se pierde la interpretación semántica de la clasificación. Todos los métodos podrían ser predictivos, en este caso.
- La clasificación de cada documento implica un proceso de tranformación previo. *Es necesario utilizar la representación del documento en el espacio de los factores*

Procesos de clasificación binarios

- Si hay H clases se desarrollan H procesos de clasificación tomando una clase cada vez como ejemplos positivos y el resto como negativos.
- Cada vez se cambia el diccionario del proceso considerando los términos más relevantes de la clase que estamos estudiando.

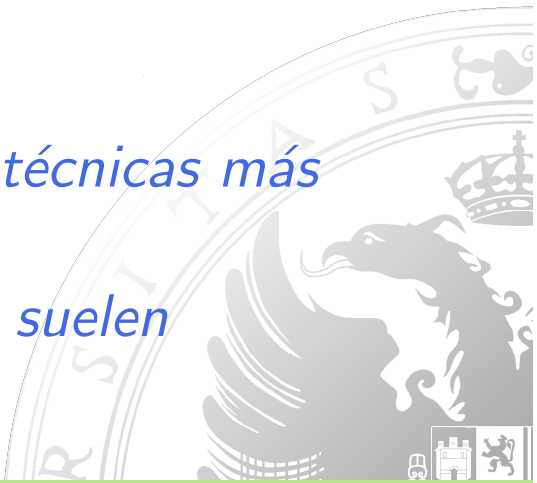


Procesos de clasificación binarios

- El proceso puede conducir a que cada documento clasificado pertenezca más de una clase:
  - Se mantiene la pertenencia a varias clases (búsqueda de tópicos)
  - Se asigna de alguna manera a la clase "ganadora" (medidas de probabilidad, medidas de distancia etc.)

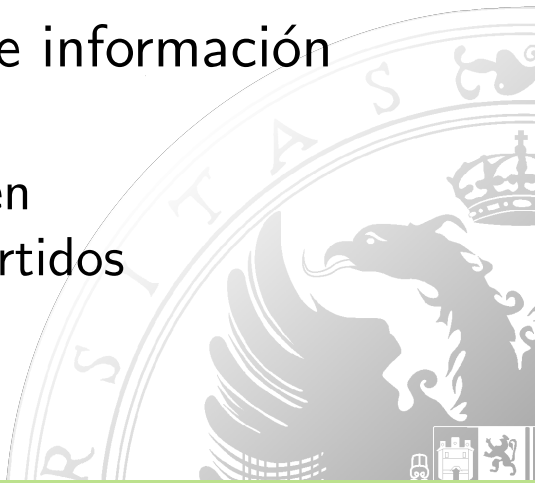
Habitualmente se trabaja con clasificaciones binarias. La adaptación de los procesos de clasificación tienen en cuenta este hecho

*Los procesos de clasificación de textos utilizan las técnicas más habituales tomando de partida el modelo vectorial documentos/términos. Según el tipo de técnica se suelen utilizar el modelo presencia/ausencia o el tf*idf*



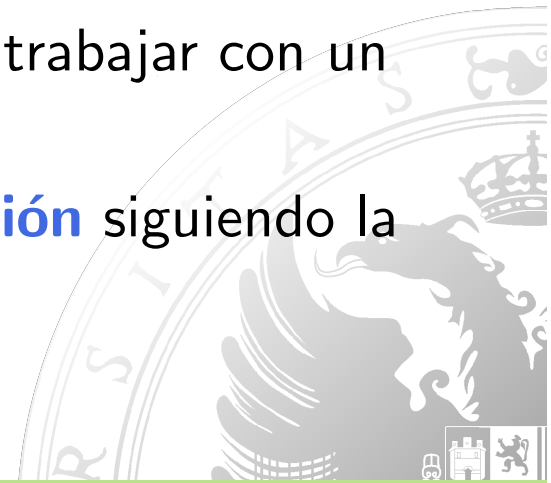
Uso de la técnica KNN

- En general se utilizan medidas de similaridad en lugar de distancias:
  - Si es un modelo binario medidas de semejanza
  - Si es un modelo tf*idf medida del coseno.
- El principal problema de aplicación es computacional: para cada documento hay que calcular la similaridad con todos.
- Solución: aplicación de técnicas de recuperación de información
  - Todo documento a clasificar es una "consulta"
  - Se obtienen los k-documentos que mejor la satisfacen
  - El proceso mejora mucho con el uso del índices invertidos (término/documentos) con los términos ordenados



Uso de técnicas basadas en reglas

- Al ser métodos explicativos están orientados más obtener las reglas que describen las clases que a clasificar.
- Habitualmente utilizan datos binarios
- Se debe intentar trabajar con reglas simples que impidan el sobreaprendizaje. Ello implica el uso de técnicas de poda.
- Datos experimentales indican que no es necesario trabajar con un gran número de términos de partida
- Tambien es posible trabajar con **árboles de decisión** siguiendo la misma filosofía anterior.



Uso de Técnicas Probabilísticas

Método Naïve Bayes

Aquí predecimos una clase binaria con N variable aleatorias binarias $(t_1, ..t_N)$, o bien ua varaible binaria N-dimensional $\bar{t}$. Las expresiones de la probabilidad condicionada, llevan a que:

$$Pr(C|\bar{t} = \bar{x}) = \frac{1}{Pr(\bar{t} = \bar{x})} exp(\sum_j w_j x_j + b)$$

donde w_i y b se pueden calcular a partir de las probabilidades estimadas . $Pr(t_j = 1|C)$, $Pr(C)$ etc. Hay dos modelos de cálculo: Bernouilli y multinomial

Uso de Técnicas Probabilísticas

Regresión Logística

Puesto que estamos clasificando de forma binaria se puede usar tf*idf y trabajar con modelos de regresión logística.Los coeficientes de la regresión logística nos pueden dar una idea del peso de los términos en la clasificación. Sobre todo si se hace por pasos.

Modelos Lineales generales.

Para cada documento D a clasificar se calcula su puntuación para pertenecer a una clase dada de forma lineal:

$$punt(D) = \sum_j w_j x_j + b$$

x_j es 0 o 1 en el documento y w_j son los "pesos de cada término en la clase". El proceso de clasificación del documentos es muy rápido con este método:

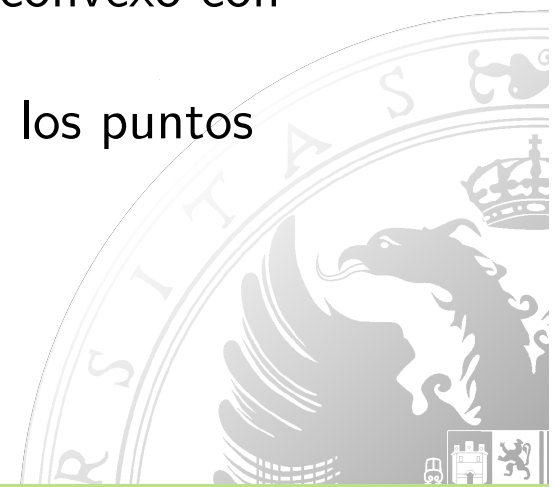
1. Se ordenan los términos junto con sus pesos (lista invertida)
2. Para cada D se recorre la lista sumando los pesos de los términos que hay en D, $sum(D)$
3. ¿$sum(D) > \alpha > 0$? se asigna el documento a la clase.

Clasificación de textos

Aplicación de técnicas de clasificación a textos

Modelos Lineales generales.

- El proceso de clasificación es el más rápido posible.
- Se pueden extender las listas de pesos utilizando, sinónimos, n-gramas etc.
- Problema: aprender los pesos
  - La optimización directa conduce a un porblema no convexo con mínimos locales.
  - Se trabaja con encontrar el hiperplano separador de los puntos positivos y negativos
  - Se utiliza normalmente una SVM



Reglas de asociación y TM

Un ejemplo interesante

- T: *términos en un documento*
- D: *conjunto de textos*
- $\{amor, dolor\} \Rightarrow \{muerte\}$

Idea básica

Un conjunto de documentos con sus términos asociados se pueden ver como una base de datos transaccional:

items\variables	t_1	t_2	$\cdots\cdots$	t_N
d_1	1	0	$\cdots\cdots$	1
$\vdots$	$\vdots$	$\vdots$	$\cdots\cdots$	$\vdots$
d_M	0	0	$\cdots\cdots$	1

Reglas de asociación y TM

Asociaciones y co-ocurrencias entre características de un texto

Formulación

◦ T =t1,t2,...,tn conjunto de términos

◦ D=d1,d2,.....,dm conjunto de documentos indexado mediante términos

Cada documento di genera un subconjunto $di(T) \subseteq T$.

Sea $U \subseteq T$. El conjunto de todos los documentos d de D tales que $U \subseteq d(T)$ se denomina *conjunto de recubrimiento* de U y se nota [U].

En su forma mas simple una regla de asociación es una implicación de la forma:

$$U \Rightarrow u \text{ con } u \subseteq T - U$$

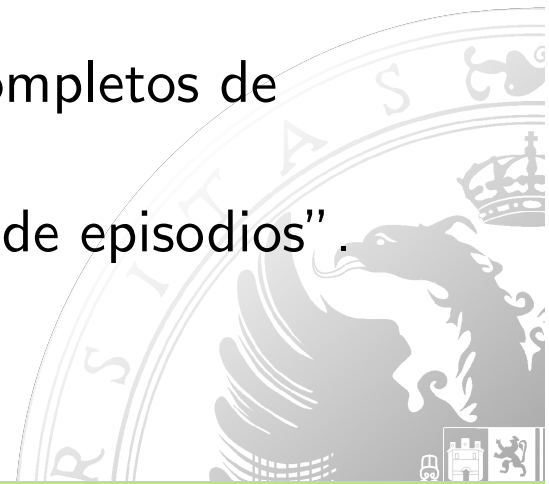
Reglas de asociación y TM

Existen diversas aproximaciones al problema dependiendo de cómo se defina T y d(T):

- T es el conjunto de todos los términos en los documentos y d(T) es la representación de documento t como una "bolsa" de palabras. Este enfoque es poco eficiente y computacionalmente costoso
- T es un conjunto de palabras clave en los documentos y d(T) es la indexación del documento T mediante un conjunto de palabras clave.
  En este enfoque el problema es obtener T. Pueden emplearse técnicas de coocurrencias basadas en el procesamiento de lenguaje natural

Reglas de asociación y TM

- Los elementos de T se agrupan según una clasificación taxonómica. (tesaurus)
  Este enfoque permite descubrir asociaciones con diferente nivel de Granularidad
- Los elementos de T son frases o sentencias completas.
  En este caso se trata de descubrir co-ocurrencias de sentencias en los documentos de la colección.
- Los elementos de T son "episodios" (conjuntos completos de características vectoriales).
  En este caso se obtienen las denominadas "reglas de episodios".



Extracción de semántica en atributos de textos cortos

Problema

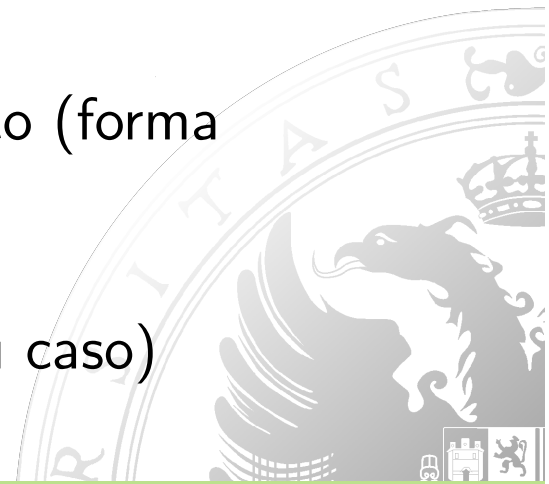
Existen atributos en bases de datos (relacionales o no) que corresponden a textos cortos con una semántica restringida.

Ejemplos:

- Bases de Datos Médicas (diagnósticos, descripción de intervenciones etc.)
- Campos de " observaciones" en diversas situaciones (encuestas, valoraciones de expertos etc.)
- Abstract de documentos

Se desea obtener una representación de dicho atributo (forma intemedia) tal que que:

- recoja la semántica
- permita la consulta como un modelo ROO (en su caso)
- permita realizar TM sobre dicha estructura



Extracción de semántica en atributos de textos cortos

Ideas básicas

- Una vez "limpios" los textos, los conjuntos/secuencias de palabras que se repiten en un número suficiente de tuplas constituyen las "frases " o "conceptos" del sistema.
- Es posible que ciertas "frases" se repitan de forma incompleta, correspondiendo a conceptos menos especializados, de forma que tenemos un retículo de subconjuntos de términos.
- Esto nos conduce al concepto de "itemset(seq) frecuente " en la base de datos transaccional de los textos. Recordemos que todo itemset frecuente tiene la propiedad "a priori"

Luego

Si obtenemos los itemset frecuentes de la base de textos:
- Los itemset maximales nos dan las frases completas
- El subretículo nos da la estructura semántica total de la base de datos.

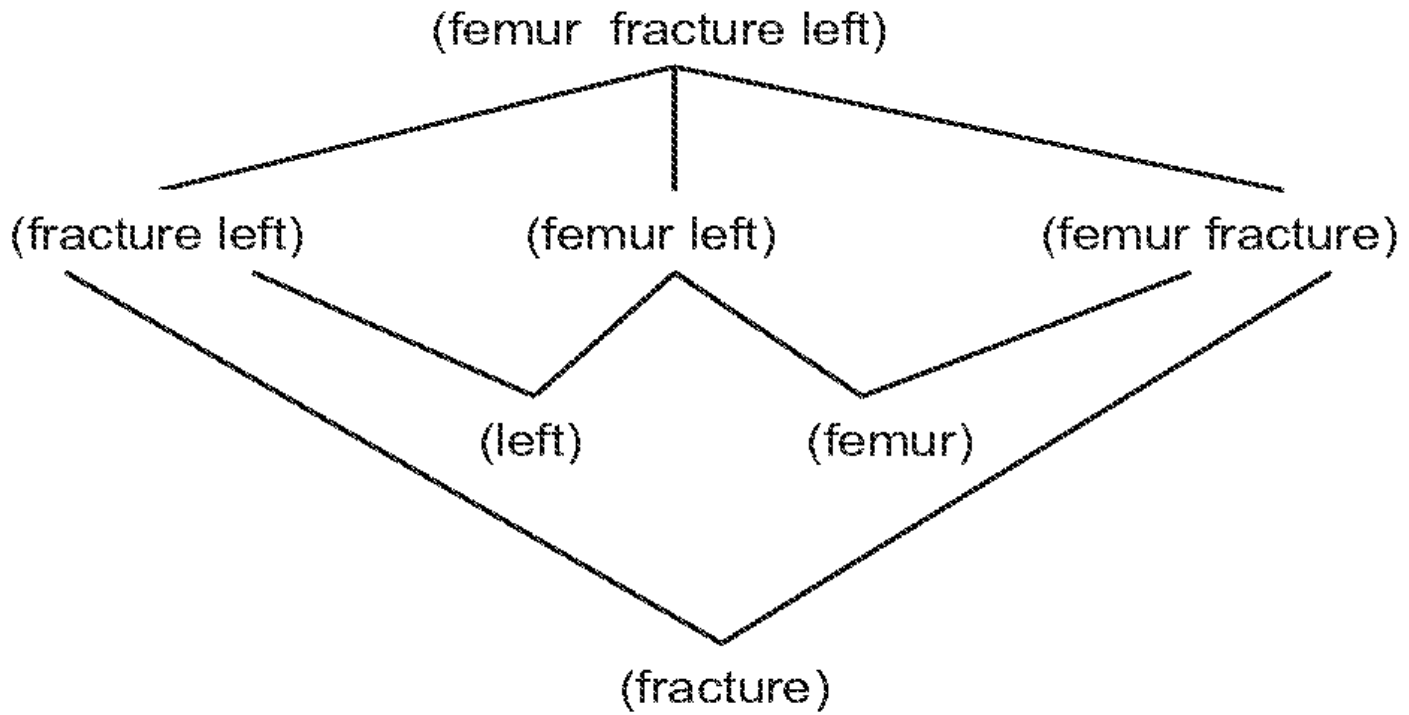
Extracción de semántica en atributos de textos cortos

Ejemplo: base de datos de urgencias

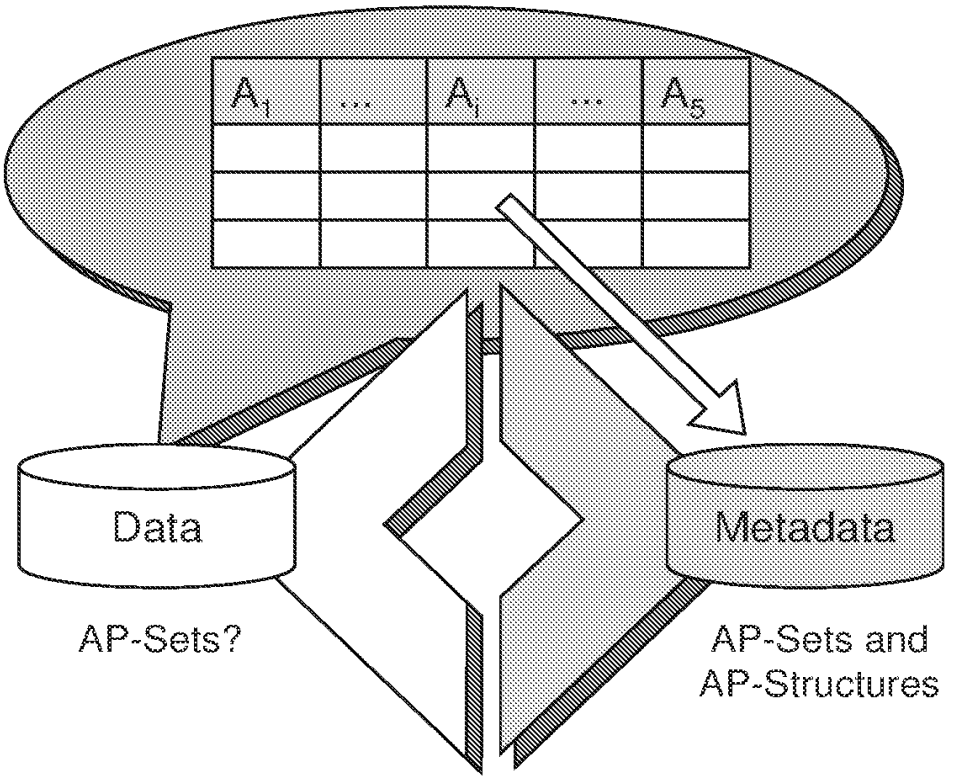
Field 1	Field 2	Field 3	...	Support
right	fracture			4.26
right	femur			1.826
back	sebaceous	cyst		1.304
back	cyst			1.304
right	femur	fracture		1.173
femur	left			1.043
femur	fracture	left		1.043
.	.			.
.	.			.
.	.			.

Extracción de semantica en atributos de textos cortos

Ejemplo: base de datos de urgencias



Extracción de semantica en atributos de textos cortos



Extracción de semantica en atributos de textos cortos

◇ La semántica global de un conjunto de textos cortos se recoje en una estructura derivada de su conjunto de itemset frecuentes
◇ El valor de cada tupla según texto que contiene es una subestructure de la global
◇ Estan definidas propiedades y operaciones de las ap-structures que permiten:

- Obtener el valor de cada tupla en el atributo como subestructura
- Realizar consultas aproximadas utilizando esta nueva estructura de datos

◇ Esta estructura se implementa como un TDA en una BDROO
◇ Hemos desarrollado TM con estas estructuras, estudiando modelos de :
- Data Warehousing
- Obtencion de itemset ordenados y generación automática de tag-clouds