# Representación del Texto

# Índice

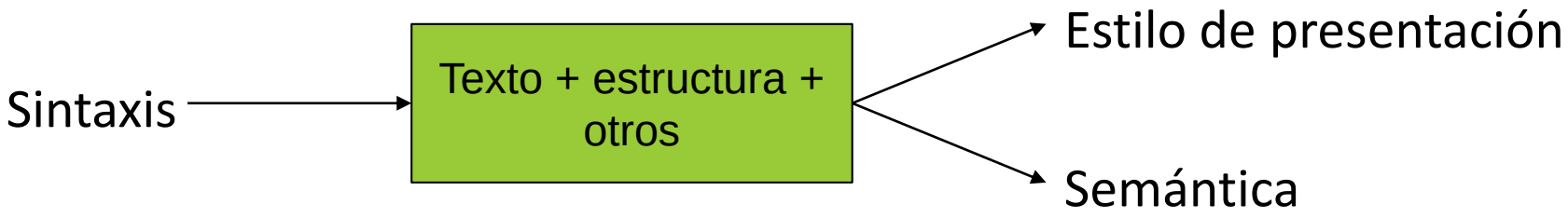
- Introducción
- Propiedades del texto
- Preprocesamiento de documentos
- Organización de documentos

# Introducción

# Introducción

- El texto ha sido el principal protocolo que los humanos han utilizado para transferir conocimiento entre generaciones
- Se han utilizado diferentes superficies de escritura: piedra, madera, pieles de animales, papiros, papel de arroz, etc.
- Actualmente, el texto puede encontrarse en papel o en formato digital
- Debido a los distintos soportes y formatos, no es fácil definir la unidad de texto
  - Definición unidad de texto: documento
- Un documento está compuesto por: sintaxis, estructura, semántica y estilo de presentación

# Introducción



# Introducción

- La sintaxis de un documento puede expresar:
  - estructura,
  - estilo de presentación,
  - semántica o
  - acciones externas
- La sintaxis puede aparecer implícita en el contenido o puede expresarse de forma explícita en un lenguaje declarativo o en un lenguaje de programación
  - Los editores de texto suelen ser declarativos
  - TeX es un lenguaje de composición tipográfica
- Tipos de lenguajes de sintaxis:
  - Propietarios y específicos
  - Abiertos y genéricos

# Introducción

- Muchos documentos tienen un formato específico
- Existe una vertiente para separar el formato del texto representado:
  - TeX / LaTeX
  - RTF
- En la mayoría de casos, el estilo lo define el autor
  - El lector suele tener alguna opción de modificar el estilo de presentación
- Las consultas de un motor de búsqueda también deben interpretarse como un documento.
  - Sus características las hacen diferentes de un texto corriente, por lo que interpretarlas bien es de vital importancia

# Propiedades del texto

# Propiedades del texto

TEORÍA DE LA INFORMACIÓN

# Propiedades del texto
# Teoría de la información

- El texto escrito es una forma de comunicar información, por lo que siempre engloba cierto contenido semántico
- Definir formalmente la cantidad de información de un texto no es sencillo
- La distribución de símbolos en un texto se puede utilizar para medir la cantidad de información representada
  - Un texto en el que sólo aparezca un símbolo no contiene mucha información
- A cada símbolo de un texto se le puede asignar una secuencia de bits o un código

# Propiedades del texto
# Teoría de la información

- Shannon (source code theory): en un esquema de codificación perfecto, un símbolo que se espera que ocurra con una probabilidad $p$, debería de tener un código con longitud $log_2 \frac{1}{p} bits$
- El número de bits en los que un símbolo se codifica, representa la información contenida por el símbolo
- Modelo estadístico del texto:
  - Probabilidad de aparición de cada símbolo
  - Código asignado a cada símbolo
- **Entropía**: cantidad media de información por símbolo en un texto completo ($\mathrm{T} = t_1, t_2 \ldots t_n$)

$$E = \frac{1}{n} \sum_{i=1}^{n} log_2 \frac{1}{p_i}$$

# Propiedades del texto
# Teoría de la información

- La entropía se calcula a partir de las probabilidades por lo que es una propiedad del modelo, no sólo del texto
- En el caso más simple, donde el modelo siempre asigna la probabilidad $p_i$ al símbolo del alfabeto $s_i$, la entropía puede definirse como:

$$E = \sum_{s_i \in \Theta} p_i log_2 \frac{1}{p_i} = -\sum_{i=1}^{\sigma} p_i log_2 p_i$$

- Donde, $\Theta$ es el alfabeto del texto, y $\sigma = |\Theta|$
- Los símbolos del alfabeto están codificados en binario, por lo que la entropía se mide en bits.
- Ejemplo: para $\sigma = 2$,
  - $E = 1$ si cada símbolo aparece el mismo número de veces
  - $E = 0$ si sólo aparece un símbolo

# Propiedades del texto

MODELADO DEL LENGUAJE NATURAL

# Propiedades del texto Modelado del lenguaje natural

- El texto está compuesto de símbolos de un alfabeto finito:
  - Símbolos para separar palabras ➔ separadores
  - Símbolos que forman palabras
- Los símbolos no se distribuyen de forma uniforme en el texto
  - Las vocales son más frecuentes que las consonantes
  - En inglés, la letra “e” tiene la mayor frecuencia
  - En español, la letra “a” tiene la mayor frecuencia
  - En el libro del Quijote, la letra “e” tiene la mayor frecuencia

# Propiedades del texto
# Modelado del lenguaje natural

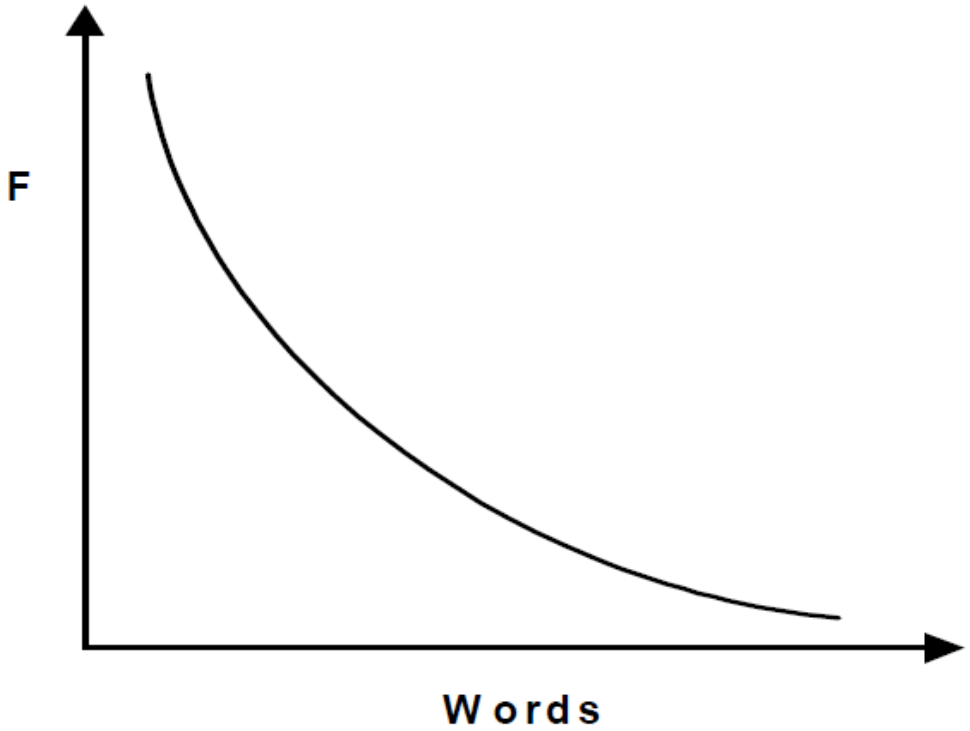
- El texto puede modelarse mediante un modelo binomial
  - Cada símbolo se genera con cierta probabilidad
  - Es un modelo simplista: existen dependencias entre los símbolos
    - Las vocales tienen una mayor probabilidad de aparición
    - Algunas consonantes no pueden aparecer juntas: la n no puede aparecer delante de una b
- La probabilidad de un símbolo depende de sus predecesores
  - Modelo de Márkov
  - Modelo de contexto finito
- El modelo puede usar $k$ letras: $modelo\ orden - k$
- Encontrar la gramática perfecta para el lenguaje natural sigue siendo un problema abierto

# Propiedades del texto
# Modelado del lenguaje natural

- La **distribución de las palabras** dentro de un texto puede aproximarse utilizando la **Ley de Zipf**
- La frecuencia $f_i$ de la palabra i-ésima más frecuente, es $\frac{1}{i^{\alpha}}$ la frecuencia $f_1$ de la palabra más frecuente, donde $\alpha$ es un parámetro dependiente del texto
  - Ejemplo: el segundo elemento se repetirá aproximadamente con una frecuencia $^1/_2$ de la del primer, el tercer elemento con una frecuencia $^1/_3$, etc.
- Los valores de $\alpha$ que mejor se ajustan a la realidad están entre 1,5 y 2
- La distribución de las palabras no está balanceada
  - Las palabras que son muy frecuentes pueden descartarse en la indexación ➔ palabras vacías (*stop words*)

# Propiedades del texto
# Modelado del lenguaje natural

- Distribución de frecuencias de un término en un documento




Fuente: Modern Information Retrieval 2nd Edition. Ricardo Baeza-Yates

# Propiedades del texto
# Modelado del lenguaje natural

- Distribución de las palabras en los documentos de una colección:
  - Cada término aparece el mismo número de veces ➔ simplista
  - El modelo que mejor se ajusta es una distribución binomial negativa

$$F(k) = \binom{a+k-1}{k} p^k (1-p)^{-\alpha-k}$$

- El número de palabras distintas en un documento forman el vocabulario
- La predicción de su crecimiento puede modelarse mediante la **Ley de Heaps**

$$V = Kn^{\beta} = O(n^{\beta})$$

# Propiedades del texto

SIMILITUD DEL TEXTO

# Propiedades del texto
# Similitud del texto

- La similitud se mide a través de una *función de distancia*
- Propiedades de las funciones de distancia:
  - Simetría: $distancia(a, b) = distancia(b, a)$
  - Desigualdad triangular:

$$distancia(a, c) \leq distancia(a, b) + distancia(b, c)$$

- **Distancia de Hamming**: dadas dos cadenas de caracteres de la misma longitud, podemos definir la distancia como el número de posiciones que tienen caracteres diferentes
  - La distancia será 0 si ambas cadenas son iguales
- **Distancia de edición o distancia de Levenshtein**: mínimo número de inserciones, eliminación o substitución de caracteres necesarios para que dos cadenas sean iguales

# Propiedades del texto
# Similitud del texto

- **Sub-secuencia común más larga (LCS)**:
  - Eliminación de los caracteres que no se repiten en dos cadenas
  - El número de caracteres restantes es la LCS
- **Similitud entre documentos**:
  - Secuencia de líneas en común entre dos documentos
  - Extraer huellas (cualquier pieza de texto significativa)
  - Distancia basada en coseno (modelo espacio vectorial)
  - Distancia basada en parecido

$$R(d_i, d_j) = \frac{|W(d_i) \cap W(d_j)|}{|W(d_i) \cup W(d_j)|}$$

# Preprocesamiento de documentos

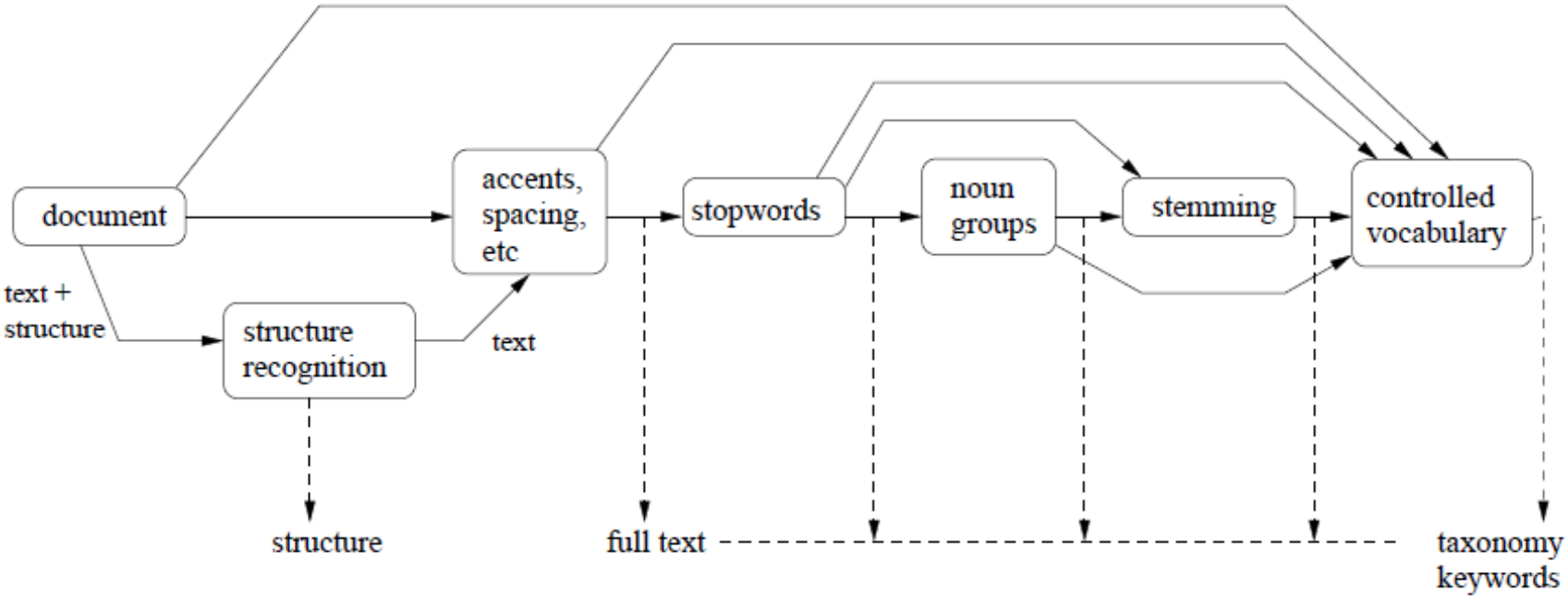
# Preprocesamiento de documentos

- El preprocesamiento de documentos puede dividirse en 5 operaciones o transformaciones:
  1. Análisis léxico del texto con el objeto de tratar:
     - Dígitos: 2015, 120, etc.
     - Guiones: saca-corchos
     - Signos de puntación: , ? ¿ ! …
     - Acentos
     - Mayúsculas y minúsculas
  2. Eliminación de las palabras vacías para filtrar aquellas palabras con valores de discriminación muy bajos para un sistema RI

# Preprocesamiento de documentos

3. Stemming (lematización) de las palabras restantes para eliminar los afijos (prefijos y sufijos)
   - Permite la recuperación de documentos que contengan variaciones sintácticas de los términos de la consulta
4. Selección de los términos/lemas (palabras, grupos de palabras, etc.) que deben emplearse como términos de indexación
   - Suelen elegirse dependiendo de la naturaleza sintáctica de la palabra
   - Los sustantivos se suelen elegir de forma preferente antes que adjetivos, adverbios o verbos
5. Construcción de categorías de términos (tesauros). Extracción de la estructura contenida en los documentos para permitir la expansión de la consulta

# Preprocesamiento de documentos




Fuente: Modern Information Retrieval 2nd Edition. Ricardo Baeza-Yates

# Preprocesamiento de documentos

ANÁLISIS LÉXICO

# Preprocesamiento de documentos Análisis Léxico

- El análisis léxico es el proceso de convertir un flujo de caracteres (el texto de un documento) en un flujo de palabras (las palabras candidatas a ser elegidas como términos de indexación)
- **Principal objetivo**: identificación de palabras en los documentos
- ¿Basta con el reconocimiento de espacios, tabulaciones y demás? !No es tan sencillo!
- Aspectos a tener en cuenta:
  - Números
  - Guiones
  - Signos de puntuación
  - Mayúsculas y minúsculas

# Preprocesamiento de documentos Análisis Léxico

- Los números no suelen ser buenos términos de indexación
  - Sin un contexto adecuado, suelen tener un significado vago
- Ejemplo: un usuario quiere saber el número de nacimientos en Roma durante el periodo 2000-2010
  - Términos de indexación: {nacimientos, Roma, 2000, 2010}
- Problema: se recuperarían todos los documentos que contengan el número 2000 y 2010
  - Cantidades monetarias
  - Velocidad
  - Temperatura
  - Etc.

# Preprocesamiento de documentos Análisis Léxico

- Casos especiales:
  - Números mezclados con letras: 5000A.C.
  - Tarjetas de crédito: conjunto de 16 dígitos
  - Números de cuentas: conjunto de 20 dígitos
  - Números de teléfonos: conjunto de 9 dígitos
  - DNI: 12345678-A
- Estos casos especiales deben tratarse con cuidado ya que son buenos candidatos a términos de indexación
  - Utilización de expresiones regulares antes de descartar
  - Las fechas pueden expresarse en diferentes formatos ➔ normalizar
  - Los números pueden expresarse con dígitos o texto

# Preprocesamiento de documentos Análisis Léxico

- Los guiones introducen una nueva problemática
  - Si el guion es un signo de puntuación deberá eliminarse
  - Si el guion forma parte de una palabra tendrá que decidirse que se hace con él
    - En algunos documentos, las palabras se dividen con guiones para que puedan entrar en una misma línea.
    - Algunas palabras pueden escribirse con guion o sin el
- Los signos de puntuación suelen eliminarse sin mayor problema
  - 1,500A.C ➔ 1500AC 
  - System.out.println(“Hola mundo”); ➔ SystemoutprintlnHolamundo 

# Preprocesamiento de documentos Análisis Léxico

- Mayúsculas y minúsculas
  - No supone un gran inconveniente
  - Todo el texto se suele convertir a minúsculas
- Casos especiales
  - La semántica puede perderse
  - Banco != banco

# Preprocesamiento de documentos

## ELIMINACIÓN DE PALABRAS VACÍAS

# Preprocesamiento de documentos Eliminación de palabras vacías

- Las palabras que son demasiado frecuentes no son adecuadas como discriminantes
  - Una palabra que aparezca en **más del 80%** de los documentos no es un buen término de indexación ➔ **Palabra vacía** (stop word)
- La eliminación de las palabras vacías nos permite reducir el tamaño de la estructura de indexación de forma considerable
  - >= 40% de reducción
- Algunos verbos, adverbios y adjetivos se suelen eliminar siempre
- Listas de palabras vacías
  - http://snowballstem.org
  - http://www.ranks.nl/stopwords/spanish

# Preprocesamiento de documentos Eliminación de palabras vacías

- Problemas
  - La eliminación de las palabras vacías puede tener como consecuencia que algunas búsquedas no tengan resultado:
    - To **be** or not to **be**
- Algunos buscadores indexan todo texto (full text index)
- La lista de palabras vacías puede construirse utilizando técnicas estadísticas
- La lista de palabras vacías dependen del contexto de la colección

# Preprocesamiento de documentos

LEMATIZACIÓN (STEMMING)

# Preprocesamiento de documentos Lematización (stemming)

- Es normal que los usuarios introduzcan un término en la consulta, pero que sólo una forma derivada se encuentre en un documento relevante
  - Plurales
  - Singulares
  - Pasados
- Una coincidencia se da entre dos palabras que sean idénticamente iguales
- Solución: sustituir las palabras por su lema (*stem* en inglés)
- El lema es lo que queda de una palabra tras eliminar los afijos:  sufijos y prefijos

# Preprocesamiento de documentos Lematización (stemming)

- Ventajas:
  - Reducción del número de variantes de una misma palabra a un mismo concepto
  - Reducción del tamaño de la estructura de indexación
    - Menos términos diferentes que indexar
- Ejemplos:
  - Conectado
  - Conectando
  - Conexión
  - Conexiones
  - Lema: conec
  - Lema: conex

# Preprocesamiento de documentos Lematización (stemming)

- La comunidad científica no tiene clara la utilización de algoritmos de stemming
  - Algunos buscadores no los incluyen
- Estrategias para la realización del stemming:
  - Basado en tablas: buscar la raíz asociada a una palabra en una tabla predefinida
    - La tabla puede ser enorme
  - Sucesores: detección de morfemas / declinaciones
  - N-grams: agrupación de palabras mediante algoritmos de clustering
  - Eliminación de afijos: eliminación del sufijo de una palabra

$$sses \rightarrow ss$$

$$s \rightarrow \emptyset$$

# Preprocesamiento de documentos Lematización (stemming)

- Algoritmos:
  - Ingles: Porter stem
  - Español, francés, portugués, rumano: romance stemmer
- Software:
  - http://snowballstem.org/
  - http://nlp.lsi.upc.edu/freeling/node/1

# Preprocesamiento de documentos

## SELECCIÓN DE LOS TÉRMINOS DE INDEXACIÓN

# Preprocesamiento de documentos Términos de indexación

- Si todo el texto se usa para la representación del texto, toda las palabras serán los términos de indexación
- En caso contrario, los términos de indexación serán un subconjunto de las palabras de los documentos
- Diferentes alternativas:
  - Taxonomía y vocabulario controlado
  - Seleccionar sólo los sustantivos
  - Seleccionar grupos de sustantivos: ingeniería informática

# Preprocesamiento de documentos

TESAURO

# Preprocesamiento de documentos Tesauro

- Un término en un tesauro denota un concepto
  - Palabras individuales
  - Conjunto de palabras (por ejemplo, sustantivo más adjetivo)
  - Frases
- Suelen ser sustantivos o verbos en gerundio que se empleen como sustantivos

# Organización de documentos

# Organización de documentos

- La organización, de entidades, objetos o cosas esta en la identidad humana.
  - Necesidad de organizar los documentos
- Con una buena organización, encontrar cierta clase o tipo de documentos se convierte en una tarea sencilla
- Sin organización, comprender y razonar sobre un gran volumen de documentos se convierte en una tarea imposible
- Esta necesidad, fue el origen de las bibliotecas
- En la actualidad existen dos grandes técnicas para la organización de documentos:
  - Taxonomías
  - Folcsonomías

# Organización de documentos

TAXONOMÍAS

# Organización de documentos Taxonomías

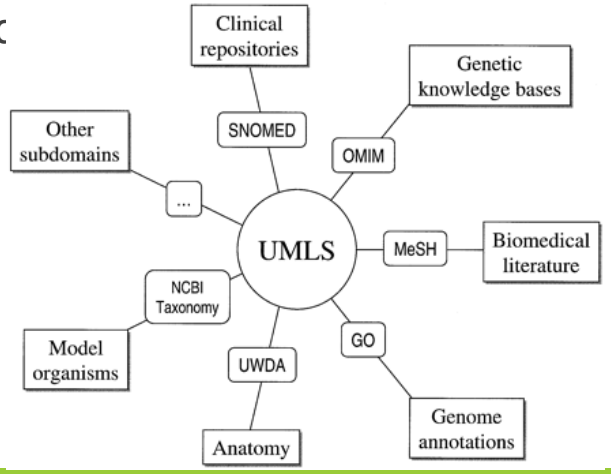
- El tipo de organización más común es la jerárquica
- Es intuitiva para los humanos
  - Los documentos, facturas, etc se suelen organizar por archivos y armarios
  - En un escritorio, pilas de documentos. Documentos similares suelen estar físicamente próximos
  - Empresas
  - Gobiernos y ministerios
  - Información dentro de un ordenador
- Permiten razonar en términos de conceptos más genéricos
- Permiten especializar, dividendo un conjunto más complejo

# Organización de documentos Taxonomías

- El tipo de organización más común es la jerárquica
- Es intuitiva para los humanos
  - Los documentos, facturas, etc se suelen organizar por archivos y armarios
  - En un escritorio, pilas de documentos. Documentos similares suelen estar físicamente próximos
  - Empresas
  - Gobiernos y ministerios
  - Información dentro de un ordenador
- Permiten razonar en términos de conceptos más genéricos
- Permiten especializar, dividendo un conjunto más complejo

# Organización de documentos Taxonomías

- Los documentos se pueden dividir en clases
- Las clases se pueden organizar de forma jerárquica utilizando, especialización, generalización y parentesco
- Este tipo de organización de se conoce como taxonomía
- Las taxonomías tienen un mayor sentido cuando se ciñen a un dominio de conocimiento
  - UMLS (Unified Medic



# Organización de documentos

FOLCSONOMÍAS

# Organización de documentos Folcsonomías

- Desventajas de las taxonomías:
  - No siempre están disponibles
  - El usuario debe conocer un vocabulario controlado
  - Los términos del vocabulario controlado no tienen por que representar adecuadamente el contenido de un documento
  - El usuario quiere poder usar sus propios términos
- La alternativa es la folcsonomías
  - Cada usuario elige los términos (llamados etiquetas)
  - El conjunto de etiquetas forma un conjunto colaborativo de descriptores de un mismo documento
  - Los usuario pueden realizar búsquedas utilizando dichas etiquetas

# Organización de documentos Folcsonomías

- Las taxonomías y las folcsonomías pueden utilizarse conjuntamente
- Se pueden utilizar técnicas automáticas para construir taxonomías a partir de folcsonomías
- La forma de representación más común para las folcsonomías es la nube de palabras

# Compresión de Texto

# Compresión de Texto

- Objetivo: encontrar la forma de representación que use el menor espacio posible
- Se basa en la identificación de regularidades
- El texto debe poder volver a su estado original
- La compresión es un tema de vital importancia:
  - Cantidad abrumadora de información
  - Reducción de costes
  - Reducción de I/O
- La desventaja es el tiempo que se pierde en la descompresión
  - Actualmente despreciable

# Compresión de Texto

- En un entorno de recuperación de información, no cualquier modelo de compresión sirve
  - Los sistemas de recuperación de información necesitan acceder a las palabras de forma aleatoria
- Se necesita que el sistema de compresión no tenga que descomprimir todo el texto para acceder a una palabra dada
- La velocidad de descompresión es más importante que la de compresión
  - Se comprime y almacena una vez
  - Se descomprime y accede multitud de veces

# Compresión de Texto

- Enfoques para la compresión de texto:
  - Estadísticos
  - Basados en diccionario
- El enfoque estadístico estima la probabilidad de aparición de un símbolo después de otro
- El enfoque basado en diccionario identifica secuencias que se repiten (llamadas frases) y que pueden ser referenciadas.
  - El conjunto de frases forman el diccionario
  - Una frase se cambia por la referencia del diccionario