UNIVERSIDAD DE GRANADA

decsai.ugr.es

Minería de Medios Sociales

DECSAI

Departamento de Ciencias de la Computación e Inteligencia Artificial

UNIVERSIDAD DE GRANADA

decsai.ugr.es

Bloque II: Minería de Texto y de la Web

DECSAI

Departamento de Ciencias de la Computación e Inteligencia Artificial

UNIVERSIDAD DE GRANADA

decsai.ugr.es

Sesión II.4 : Minería de Texto y Ontologías

DECSAI

Departamento de Ciencias de la Computación e Inteligencia Artificial

Minería de Texto

La Minería de Texto (Text Mining, TM) [Feldman, Dagan, 1995] se define como la aplicación a textos de algoritmos y métodos del campo del aprendizaje automático (Machine Learning, ML) y la estadística, con el objetivo de encontrar patrones útiles.

Para lograr este propósito es necesario pre-procesar los textos.

- Tokenización
- Eliminación de palabras vacías — *stop word removal*
- Etiquetado categoría gramatical — *part-of-speech tagging*
- Lematización — *stemming*
- Detección de Sinónimos — *synonym detection*
- Detección de Entidades — *Named Entity Recognition*
- Representación intermedia — *Bag of Words*
- Esquema de pesos — *TF-IDF, ...*

Hay tantas enfoques sobre la Minería de Texto como aplicaciones.
La minería de Texto es por definición un campo de investigación muy orientado a las aplicaciones.

Minería de Texto

La mayoría del contenido textual disponible no posee estructura alguna.

En Minería de Texto es necesario que los textos estén estructurados de algún modo para poder procesarlas de forma adecuada. Habitualmente se trabaja con técnicas orientadas a los datos:

- Recuperación de información - esquemas de pesos
- Clustering de Texto – modelos de categorización no supervisada
- Clasificación de Texto – modelos de categorización supervisada

- Procesamiento de Lenguaje Natural
  - Reconocimiento del habla – modelos probabilísticos
  - Traducción automática – modelos probabilísticos

Es posible utilizar herramientas para dotar de significado al texto:

- Representación de estructuras conceptuales
  - Tesauros
  - Ontologías

Minería de Texto y Ontologías

Utilizando Minería de Texto podemos obtener información muy útil a partir de grandes cantidades de texto.

Sin embargo, para poder obtener mejores resultados es necesario comprender la **semántica** del texto procesado.

Las **ontologías** nos permiten representar y organizar la información **semántica** de cara a poder emplearla en procesos de minería de texto.

Introducción a las Ontologías I

Una especificación explícita y formal sobre una conceptualización compartida [Gruber, 1993]

Las ontologías describen **conceptos** y **relaciones** existentes en algún dominio, de forma compartida y consensuada.
Generalmente representan una jerarquía de conceptos junto a relaciones entre conceptos que pueden directas, transitivas, reflexivas, etc.
Esta conceptualización debe ser representada de una manera **formal**, **legible** y utilizable mediante un **procesamiento automático**.

Las ontologías no sólo permiten la **estructuración** del conocimiento, sino que también permiten realizar un **razonamiento** sobre las afirmaciones que modelan.
Mediante el uso de razonadores, se puede validar una ontología o realizar tareas tales como clasificación de instancias y clases.

Introducción a las Ontologías II

Formalmente, una ontología está formada por:

- **Clases:** Son los conceptos del dominio.
- **Propiedades:** Pueden ser de dos tipos...
  - ***Relaciones***: enlaza dos clases de la ontología.
  - ***Atributos***: son las características propias de una clase.
- **Individuos**. Son las instancias concretas de una clase.
- **Axiomas**. Son restricciones impuestas a los elementos de la ontología

Introducción a la Ontologías III

Existen diferentes tipos de ontologías según el tipo de conocimiento que modelan

- **Ontología de Alto Nivel:** Describen conceptos muy generales como espacio, tiempo, eventos, que son independientes de un problema o dominio particular. Lo más razonable es tener ontologías de alto nivel comunes para grandes comunidades de usuarios.
- **Ontología de Dominio:** Describe el vocabulario asociado a un dominio genérico, especializando los conceptos introducidos en la ontología de alto nivel.
- **Ontología de Tarea:** Describe el vocabulario asociado a una actividad o tarea genérica especializando una ontología de alto nivel.
- **Ontologías de Aplicación:** Son las ontologías mas específicas. Los conceptos de estas ontologías suelen corresponderse con los roles desempeñados por las entidades de dominio cuando se realiza una cierta actividad.

Introducción a la Ontologías IV

Las Ontologías formalizan la parte **intensional** del conocimiento *(estructura)* sobre un dominio mientras la parte **extensional** *(datos)* la proporciona una Base de Conocimiento (Knowledge Base, KB), que contiene aserciones sobre las instancias de los conceptos y relaciones de la ontología.

Podríamos decir que el conocimiento en una KB se rige por una conceptualización explícita o implícita. Dicha conceptualización es la ontología.

Lenguajes de Representación de Ontologías

El World Wide Web Consortium (W3C) es un consorcio internacional compuesto por organizaciones e investigadores, cuyo objetivo es desarrollar estándares para la web.

En el caso de representación de ontologías los principales estándares son:

- RDF (Resource Description Framework)
- RDFS (RDF Schema)
- OWL/OWL2 (Web Ontology Language)

Web Ontology Language (OWL 2) I

Introducción básica
http://www.w3.org/TR/owl2-overview/

Guía de referencia rápida del lenguaje
http://www.w3.org/TR/owl2-quick-reference/

OWL 2 es un lenguaje de para la **definición de ontologías** en la Web Semántica.

Las ontologías OWL 2 representan **clases**, **propiedades**, **individuos**, y **valores de datos**.

OWL 2 puede expresarse usando **varias sintaxis**, la única reconocida como obligatoria es la RDF/XML.
Las ontologías OWL 2 pueden verse como grafos RDF.

Es posible **razonar** de forma automática sobre OWL 2.

Web Ontology Language (OWL 2) II

La **semántica** asociada a las estructuras del lenguaje OWL 2 puede asignarse según dos criterios:

- **Semántica Directa (OWL 2 DL):** Se corresponde con la Lógica Descriptiva (DL) SROIQ, pero limita el uso de algunas estructuras y propiedades, para garantizar que sea decidible.

- **Semántica RDF (OWL 2 Full):** Se aplica la semántica que se usa en RDF, tiene todo el poder expresivo de representación, pero limita las capacidades de razonamiento al no ser siempre decidible.

Web Ontology Language (OWL 2) III

Atendiendo a su capacidad expresiva se definen varios perfiles:

- **OWL 2 EL:** Apropiado para aplicaciones que usan grandes ontologías, y donde se puede sacrificar la capacidad expresiva para garantizar el rendimiento.
- **OWL 2 QL:** Apropiado para aplicaciones con ontologías ligeras (tamaño pequeño y poca complejidad) y con un alto número de instancias, que se organizan para poder usar consultas relacionales.
- **OWL 2 RL:** Apropiado para aplicaciones en las que ontologías relativamente ligeras se usan para organizar gran cantidad de individuos y es necesario operar sobre los datos como triples RDF.

Las ontologías de OWL 1, son por definición ontologías OWL 2 válidas.

Minería de Texto y Ontologías I

Atendiendo al uso que se puede hacer de las ontologías en procesos de minería de texto podemos hablar de ontologías comunes y ontologías de dominio.

- **Ontologías comunes:** Donde se representan relaciones de objetos generales. Tienen el inconveniente de que carecen en muchos casos de lenguaje específico técnico de un dominio particular.

  Un ejemplo son los diccionarios semánticos tales como WordNet.

- **Ontologías de dominio:** Representan un vocabulario y un conjunto de relaciones más reducido sobre un dominio específico. Se usan para representar vocabulario específico de un dominio concreto que no es de uso común.

  Suelen realizarse por expertos, de forma manual o semi-automática, y por tanto, su construcción es costosa.

Minería de Texto y Ontologías II

Las ontologías en Minería de Texto desempeñan diferentes roles atendiendo a su uso, de forma general podemos distinguir:

- **Ontologías como recurso semántico.**

- **Ontologías como producto de un proceso de Minería de Texto:**
  - Aprendizaje de Ontologías
  - Representación mediante Ontologías

Ontologías como Recurso Semántico

Para poder conocer el significado del texto que se procesa es necesario acudir a fuentes de información que determinen la semántica de los términos procesados.

Suggested Upper Merged Ontology (SUMO)

Yet Another Great Ontology (YAGO)

Systematized Nomenclature of Medicine (SNOMED)

Suggested Upper Merged Ontology (SUMO)

Es una de las mayores ontologías formales existentes.

- Ontología gratuita propiedad de IEEE (Diciembre 2000).
- Alrededor de 25.000 términos y 80.000 axiomas.
- Mapeada a todos los lexicon de WordNet.
- Escrita en el lenguaje SUO-KIF (Standard Upper Ontology *Knowledge Interchange Format*).
- Se emplea en investigación y aplicaciones sobre búsqueda, lingüística y razonamiento.
- Las ontologías que extienden SUMO deben estar disponibles bajo la licencia GNU General Public License.

- Web SUMO: http://www.adampease.org/OP/

Yet Another Great Ontology (YAGO)

Ontología extraída de forma automática a partir de Wikipedia, WordNet y Geonames:

- Desarrollada en el Instituto Max Planck (2008).
- Existe una versión actualizada YAGO2 (2012).
- Contiene 10 millones de entidades y 120 millones de hechos.
- YAGO está enlazada a SUMO y la Ontología de DBPedia.
- Se utilizó en el sistema Watson de IBM.
- Se distribuye con licencia Creative Commons.

Demo:
http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/demo/

Systematized Nomenclature of Medicine (SNOMED)

Es la mayor recopilación de terminología médica multilingüe del mundo.

- Contiene términos sobre anatomía, enfermedades, procedimientos, microorganismos, etc...
- Originalmente desarrollada en 1973, revisada en 1990 y en 2002 se amplió con información proporcionada por el NHS inglés.
- Las versiones previas de SNOMED dejarán de tener validez en 2017.

Explorador: http://browser.ihtsdotools.org/

Aprendizaje de Ontologías I

Consiste en la generación automática o semi-automática de ontologías utilizando técnicas de aprendizaje automático (ML) o de procesamiento de lenguaje natural (NLP).

Las primeras referencias al término (Ontology Learning), podemos encontrarlas en [Madche and Staab, 2001] donde se describe en términos de la adquisición de un modelo de un domino a través de los datos.

Cuando el aprendizaje de ontologías se realiza sobre fuentes textuales no estructuradas, es cuando hablamos de aprendizaje de ontologías a partir de texto.

Aprendizaje de Ontologías II

El proceso de aprendizaje de ontologías puede verse como un proceso de ingeniería inversa, pero presenta los siguientes inconvenientes:

- En el proceso de creación de un texto sólo refleja el dominio de conocimiento del autor de forma parcial, por lo que el proceso de ingeniería inversa como mucho podrá reconstruir parcialmente dicho modelo de conocimiento
- El conocimiento del mundo raramente se menciona de forma explícita

Aprendizaje de Ontologías III

No existe un consenso entre la comunidad investigadora en cuanto a cuales son las tareas concretas que deben realizarse en el proceso de aprendizaje.
En nuestro caso nos ceñiremos a la definición de subtareas para el desarrollo de ontologías definida en [Cimiano, 2006].

$\forall x(país(x) \rightarrow \exists y\ capital\_de(y,x) \wedge \forall z(capital\_de(z,x) \rightarrow y=z))$

disjuntos(río, montaña)

$capital\_de \leq_R localizado\_en$

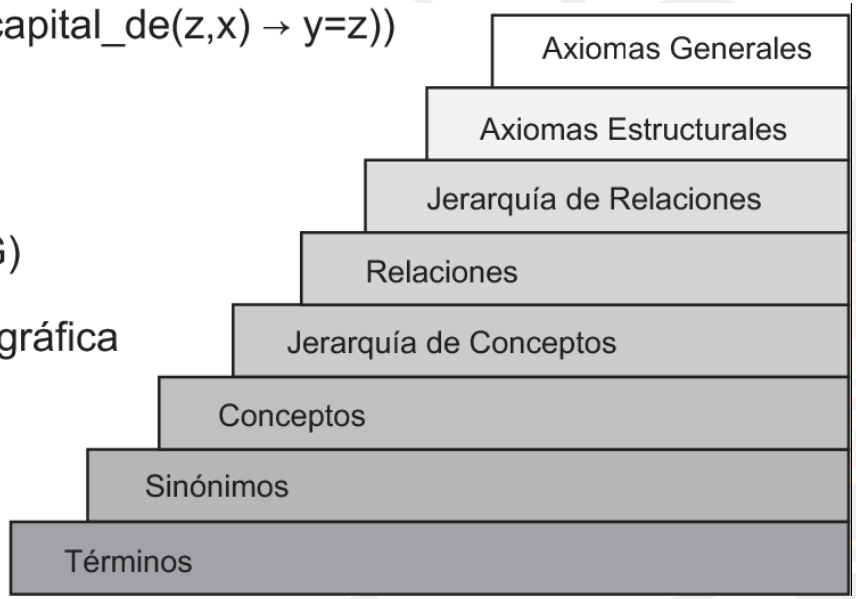
fluye_a_través(dominio:río, rango:ZG)

$capital \leq_c ciudad$, $ciudad \leq_c$ zona geográfica

$c:=$ país $:= \langle i(c), ||c||, Ref_c(c) \rangle$

[país, nación]

río, país, nación, ciudad, capital ...



Aprendizaje de Ontologías IV

Para realizar cada una de las subtareas se suelen utilizar distintas técnicas:

Términos – frecuencia, TFIDF, entropía, estadísticas de un corpus de referencia
Sinónimos – diccionarios electrónicos (implica WSD)
Conceptos - diccionarios electrónicos
Jerarquía de conceptos – clustering, patrones de Hearst
Relaciones – reglas de asociación
Jerarquía de relaciones - clustering
Axiomas estructurales
Axiomas generales

Tareas

Las ontologías en Minería de Texto, se pueden emplear para diversas tareas:

- Extracción de Información (IE)
- Recuperación de Información (IR)
- Aprendizaje de Ontologías (OL)
- Población de Ontologías (OP)
- Detección de eventos (NED)

Aplicaciones

Existen diversas aplicaciones en las que se emplean ontologías para Minería de Texto:

- Clasificación de páginas web en Directorios Web
- Agrupamiento de Documentos Médicos Multilingues
- Creación de ontologías de dominio a partir de Texto

Otras técnicas que usan Semántica

La amplia difusión que ha tenido el uso de ontologías en procesos de Minería de Texto, se debe a la importancia de contar con herramientas que permitan gestionar la semántica de los datos textuales.

Existen otras técnicas que tratan de obtener dicha semántica no a través del uso de fuentes de conocimiento externas, sino a través del descubrimiento de la semántica latente del texto, a través de técnicas estadísticas.

Entre estas técnicas destacan:

- Análisis de Semántica Lantente (Latent Semantic Analysis / **LSA**)
- Latent Dirichlet Allocation (**LDA**)

Análisis de Semántica Latente

Latent Semantic Analysis (LSA) [Deerwester et al. 1990], es una técnica algebraica de análisis factorial que premite reducir la dimensionalidad de una matriz de términos-documentos, capturando la mayor parte de la varianza de un corpus textual.

LSA parte de la hipótesis de que palabras con significados similares ocurrirán en contextos similares.

Cuando comparamos documentos, lo hacemos utilizando los términos. LSA permite comparar los documentos a un nivel más general, a nivel de concepto o características/factores (*features*).

Utilizando descomposición en valores singulares (Singular Value Decomposition SVD) podemos extraer esas características de los documentos.

Análisis de Semántica Latente

Mediante SVD la matriz de documentos-términos A se separa en 3 matrices.

$$A = U \sum V^T$$

- U : Relaciona cada término con los nuevos conceptos encontrados.
- $\sum$ o S : Matriz diagonal que contiene los valores singulares de A representados en orden descendente.
- V^T: Relaciona los documentos con los conceptos.

Ejemplo LSA

Tomamos un conjunto de documentos y sus términos correspondientes, dado que sólo se tienen en cuenta propiedades estadísticas, en el ejemplo se usan letras para representar términos.

```
d1: c a a b c b c
d2: a b c a b c c
d3: d e f f d
d4: f d e d f
```

Creamos la matriz de términos-documentos donde las filas representan términos únicos y las columnas documentos. El contenido de la celda en este caso será la frecuencia absoluta, pero podía usarse un esquema de pesos como TF-IDF.

	d1	d2	d3	d4
a	2	2	0	0
b	2	2	0	0
c	3	3	0	0
d	0	0	2	2
e	0	0	1	1
f	0	0	2	2

Ejemplo LSA: http://matpalm.com/lsa_via_svd/index.html

Ejemplo LSA

Aplicamos SVD y obtenemos las siguiente matrices:

A = U x S x Vt

	d1	d2	d3	d4
a	2	2	0	0
b	2	2	0	0
c	3	3	0	0
d	0	0	2	2
e	0	0	1	1
f	0	0	2	2

	f1	f2	f3	f4
a	0.48	0	0	0
b	0.48	0	0	0
c	0.72	0	0	0
d	0	-0.66	0	0
e	0	-0.33	0	0
f	0	-0.66	0	0

	f1	f2	f3	f4
f1	5.83	0	0	0
f2	0	4.24	0	0
f3	0	0	0	0
f4	0	0	0	0

	d1	d2	d3	d4
f1	0.70	0.38	0	0
f2	0	0	-0.70	-0.70
f3	0	0	0	0
f4	0	0	0	0

En **S** podemos ver como tenemos dos conceptos/características en f1 y f2

En **U** vemos como se relaciona cada término con los conceptos encontrados

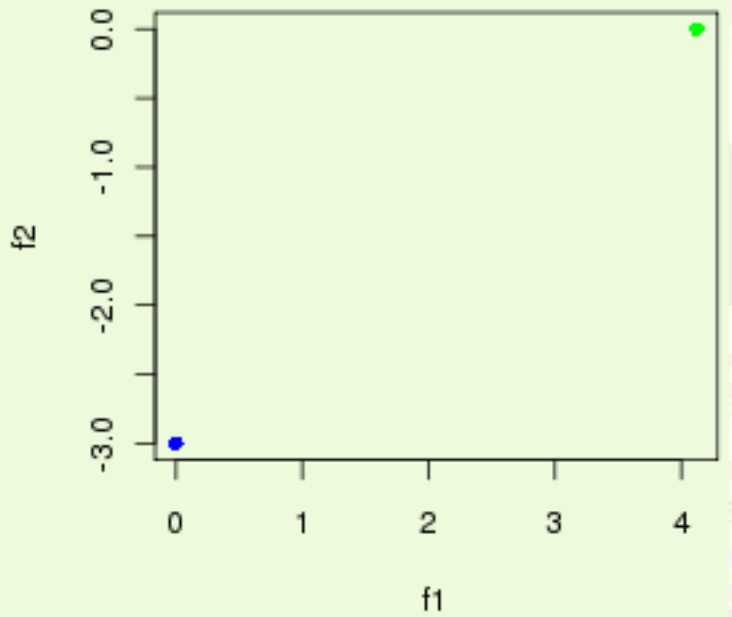
En $\mathbf{V}^T$ vemos la relación de los documentos con los conceptos.

Ejemplo LSA

Si multiplicamos S y VT obtenemos la relación entre los documentos y los conceptos.

Representando en una gráfica f1 y f2 vemos como en efecto existe una separación evidente entre los documentos d1 y d2, respecto a d3 y d4.

	f1	f2	f3	f4
d1	4.123	0.000	0.000	0.000
d2	4.123	0.000	0.000	0.000
d3	0.000	-3.000	0.000	0.000
d4	0.000	-3.000	0.000	0.000



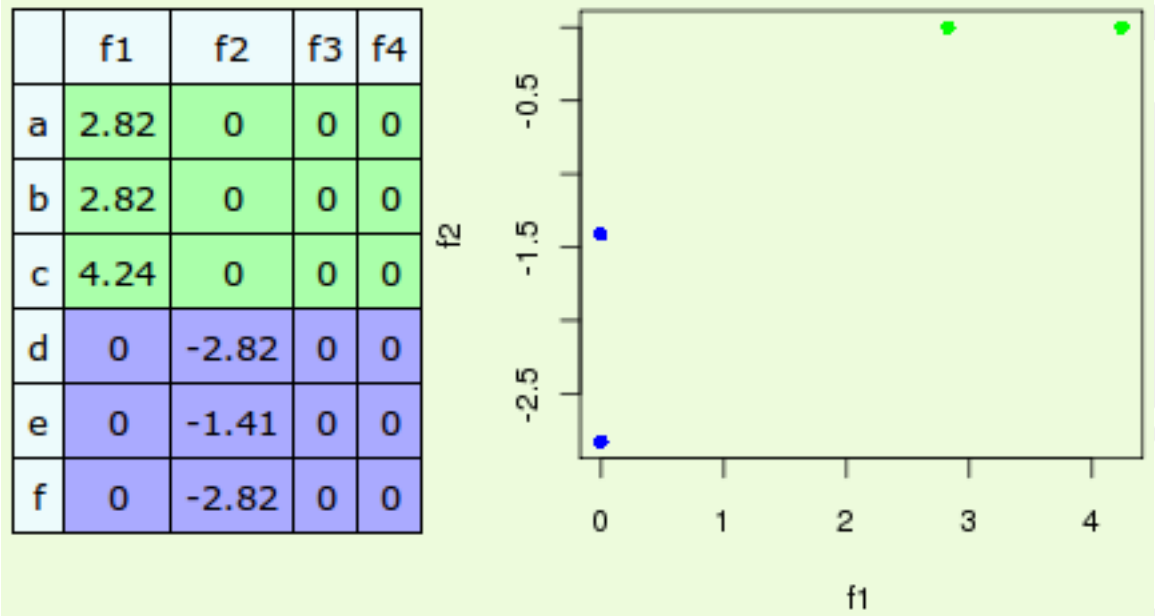
Ejemplo LSA

Si multiplicamos las matrices U y S, obtendremos la relación entre los términos y los conceptos.

Los términos *a*, *b* y *c* se alinean con el concepto 1, mientras que d, e y f lo hacen con el concepto 2.

También vemos como c tiene una asociación más fuerte con el concepto 1 que a o b (porque tiene mayor frecuencia), al igual que e tiene una menor asociación con el concepto 2, que d o e (por su menor frecuencia).

	f1	f2	f3	f4
a	2.82	0	0	0
b	2.82	0	0	0
c	4.24	0	0	0
d	0	-2.82	0	0
e	0	-1.41	0	0
f	0	-2.82	0	0



Análisis de Semántica Latente

Ventajas:

• Permite reducir la dimensionalidad de la matriz de datos

• Al ser un enfoque puramente numérico se puede emplear con cualquier idioma.

Limitaciones:

• Dada la alta dimensionalidad de los datos textuales, calcular SVD es muy costoso.

• No tiene en cuenta el orden de las palabras

• El nuevo espacio de conceptos/características es muy difícil de interpretar al ser una combinación lineal de un conjunto de palabras del espacio original.

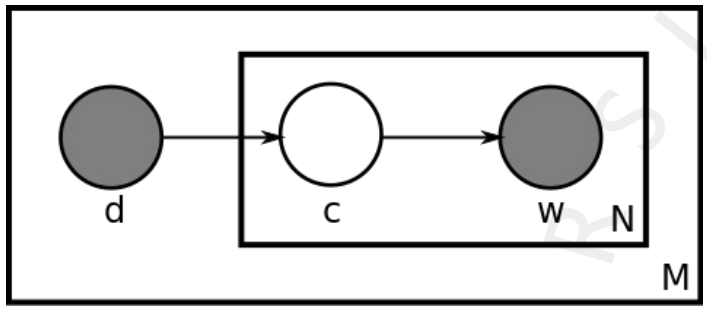
• No se puede generalizar para incluir información adicional (fechas, autores...).

Probabilistic LSA (pLSA)

También conocido como *aspect model*, es una alternativa a LSA.

Es un modelo de variable latente que asocia una variable de clase no observada *c* (aspecto/concepto), con cada documento de la colección *d* y representa cada aspecto como una distribución de palabras con una determinada probabilidad *P(w|c)*.

$$P(w,d) = \sum_c P(c)P(d|c)P(w|c) = P(d)\sum_c P(c|d)P(w|c)$$



http://commons.wikimedia.org/wiki/File:Plsi_1.svg

d es la variable del documento, c es un concepto al que pertenece una palabra obtenida mediante la distribución P(c|d), w es una palabra obtenida de la distribución de palabras del concepto al que pertenece esa palabra P(w|c). Las variables d y son variables observadas, el concepto c es una variable latente.

Latent Dirichlet Allocation (LDA)

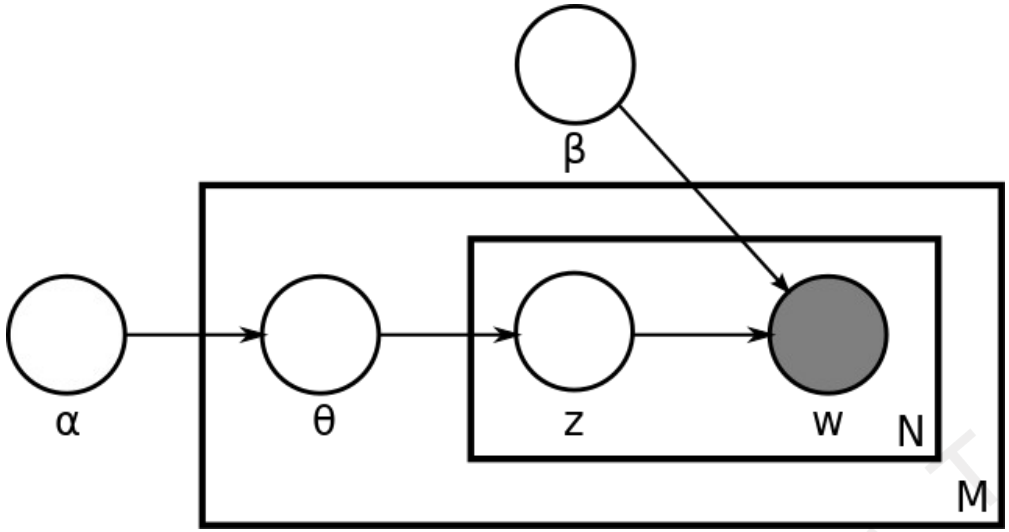
Es un modelo generador probabilístico de un corpus de documentos. La idea es que que cada documento está representado por una mezcla de temas, donde cada tema es una variable latente caracterizada por una distribución sobre un vocabulario fijo de palabras.

La completitud del proceso generador para un documento se consigue considerando a priori una distribución de Dirichlet en el documento sobre los temas y en los temas sobre las palabras.

La estructura del modelo LDA permite la interacción de las palabras observadas en documentos con las distribuciones estructuradas de un modelo de variables oculto.

Este método se ha aplicado para encontrar estructuras útiles en distintos tipos de documentos, tales como emails, literatura científica, libros en librerías digitales y archivos de noticias.

Latent Dirichlet Allocation (LDA)



http://commons.wikimedia.org/wiki/File:Latent_Dirichlet_allocation.svg

α : Es el parámetro de la distribución Dirichlet para las distribuciones de temas por documento.
β: Es el parámetro de la distribución Dirichlet para las distribuciones de palabras por tema.
θ: Es la distribución de temas por documento i
z: Es el tema para la palabra w_{ij} en el documento i
w_{ij}: Es una palabra concreta

Solo w_{ij} es una variable observable, el resto son ocultas.

Online LDA

Existe una variante Online de LDA para analizar flujos de texto [AlSumait et al. 2008].

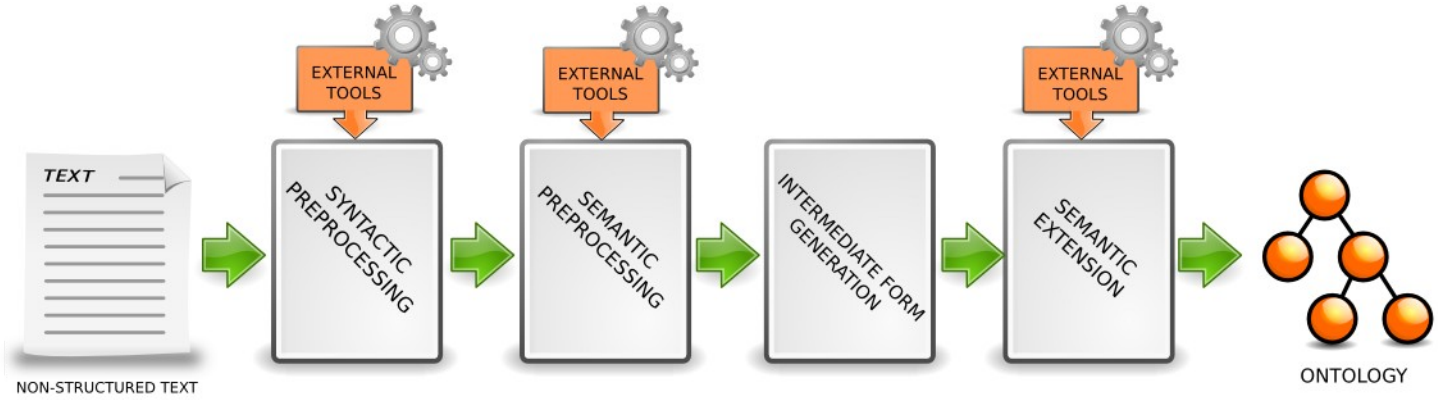
El modelo OLDA considera la ordenación temporal de la información y asume que los documentos van llegando en intervalos de tiempo discretos.

En cada intervalo de tiempo de un tamaño determinado (hora, día, año...) un flujo de documentos de tamaño variable se recibe para ser procesado.

Un documento recibido se representa como un vector de palabras.

Entonces se usa LDA para modelar los documentos recibidos. El modelo generado en un instante determinado, se usa como distribución a priori en el siguiente intervalo de tiempo, cuando lleguen nuevos documentos.

Ejemplo: Generación de Ontologías I

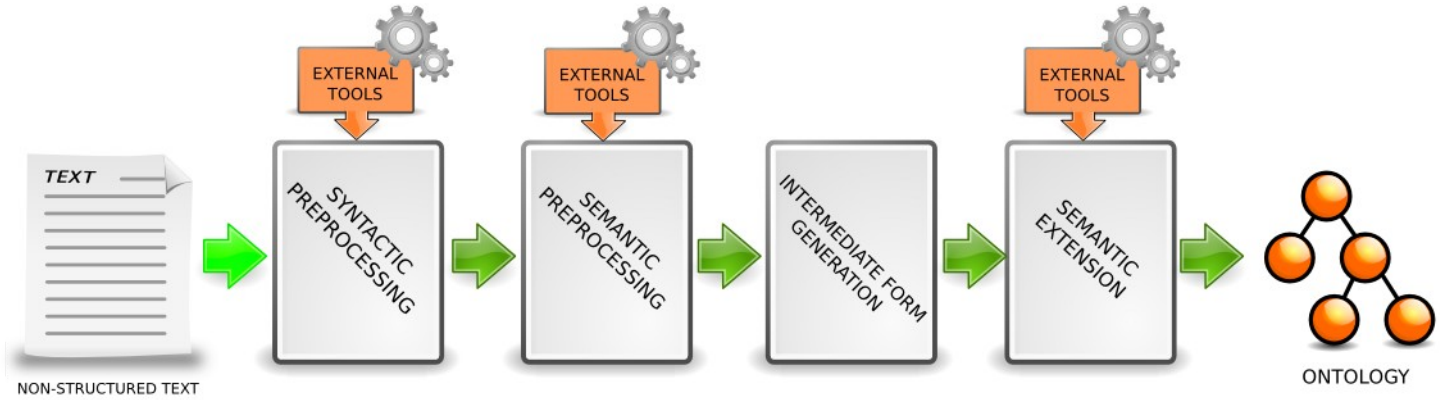


16616 FREDDI: A FUZZY RELATIONAL DEDUCTIVE DATABASE INTERFACE
21186 DATA SUMMARIZATION IN RELATIONAL DATABASES THROUGH FUZZY DEPENDENCIES
164017 GEFRED - A GENERALIZED-MODEL OF FUZZY RELATIONAL DATABASES

El objetivo es la generación de una ontología que represente los principales temas de los que tratan los textos.

Se emplea una metodología genérica que puede instanciarse con distintas herramientas en diferentes etapas.

Ejemplo: Generación de Ontologías II

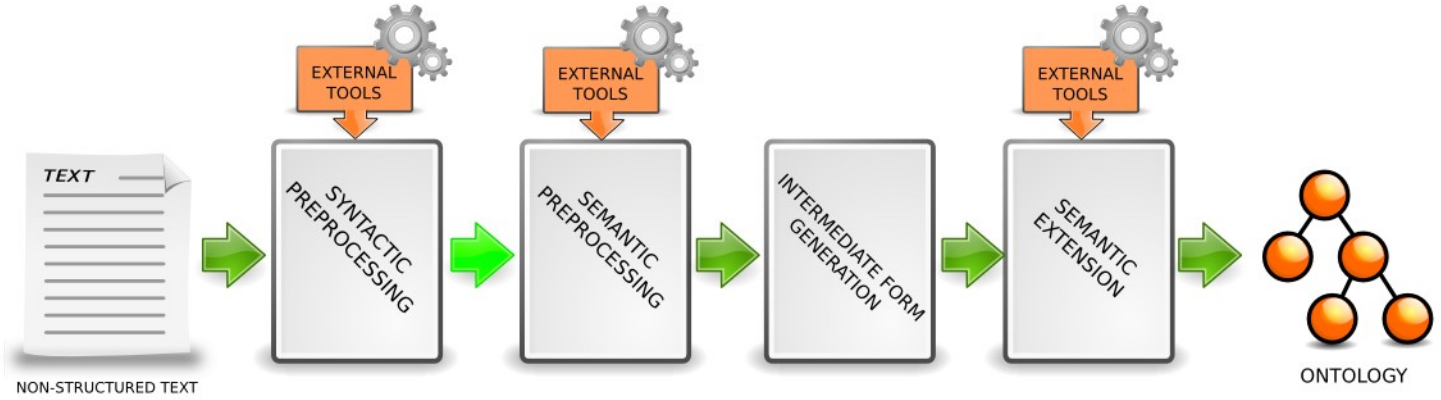


16616 FREDDI: A FUZZY RELATIONAL DEDUCTIVE DATABASE INTERFACE
21186 DATA SUMMARIZATION IN RELATIONAL DATABASES THROUGH FUZZY DEPENDENCIES
164017 GEFRED - A GENERALIZED-MODEL OF FUZZY RELATIONAL DATABASES

Preprocesamiento sintáctico:

- Tokenización
- Eliminación de palabras vacías
- Etc...

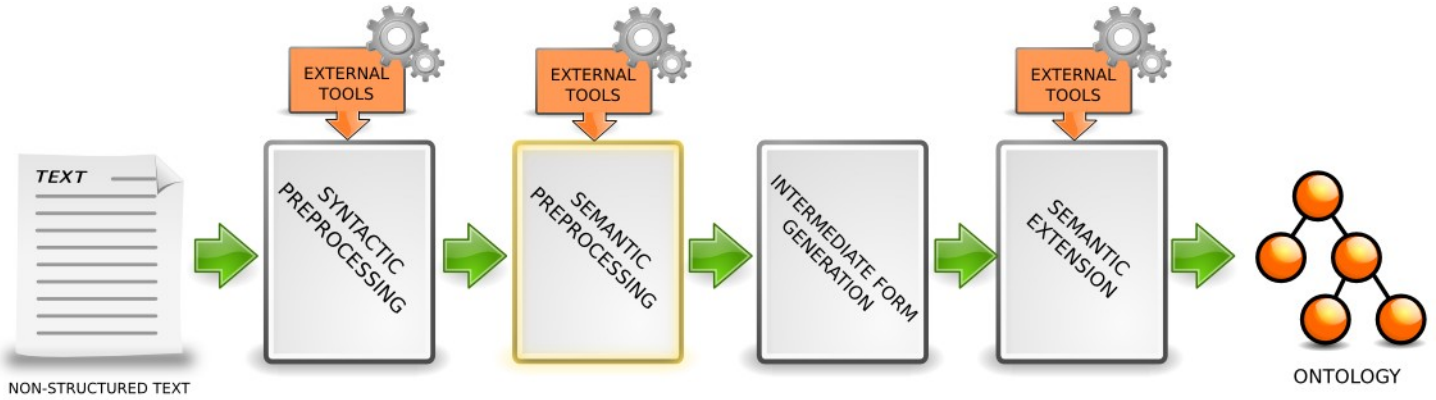
Ejemplo: Generación de Ontologías III



Preprocesamiento semántico:

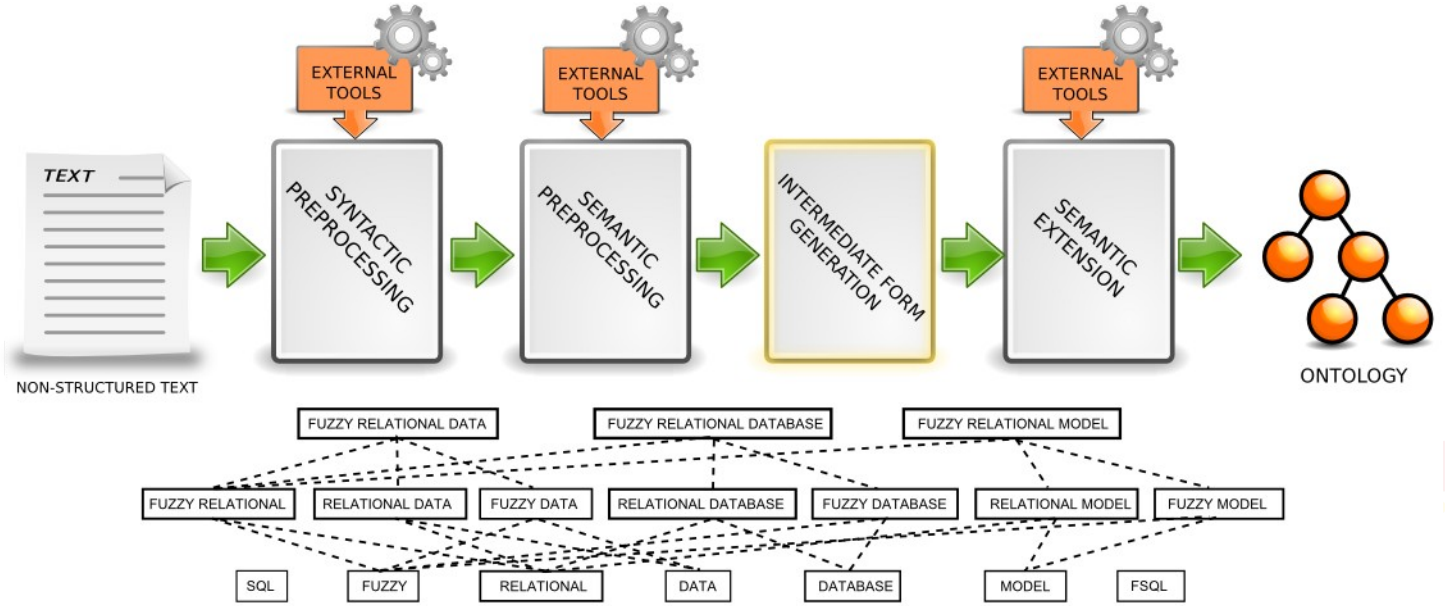
- **Unifica** los términos
- Determina el conjunto del sinónimos del término
  - Categoría gramatical (POS)
  - Desambiguación (WSD)

Ejemplo: Generación de Ontologías IV



- Determinar el **representante canónico** de cada conjunto de sinónimos
- Sustituir los términos por el representante canónico del conjunto de sinónimos.

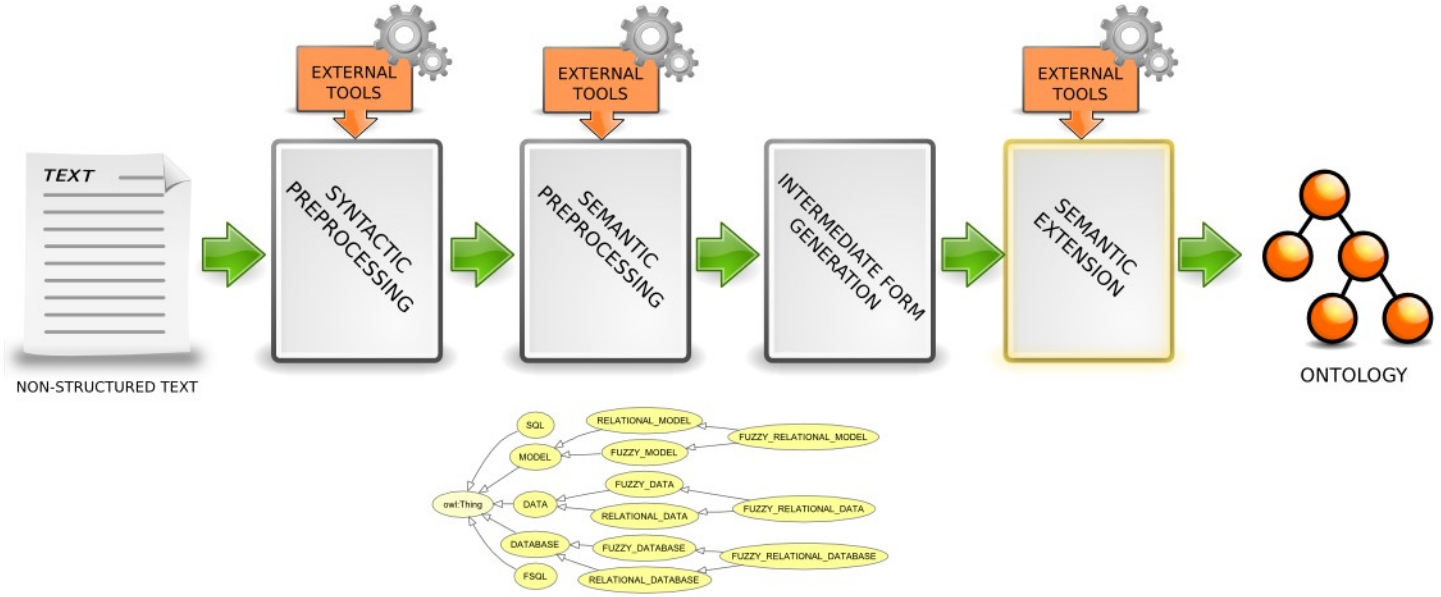
Ejemplo: Generación de Ontologías V



Se genera la forma intermedia:

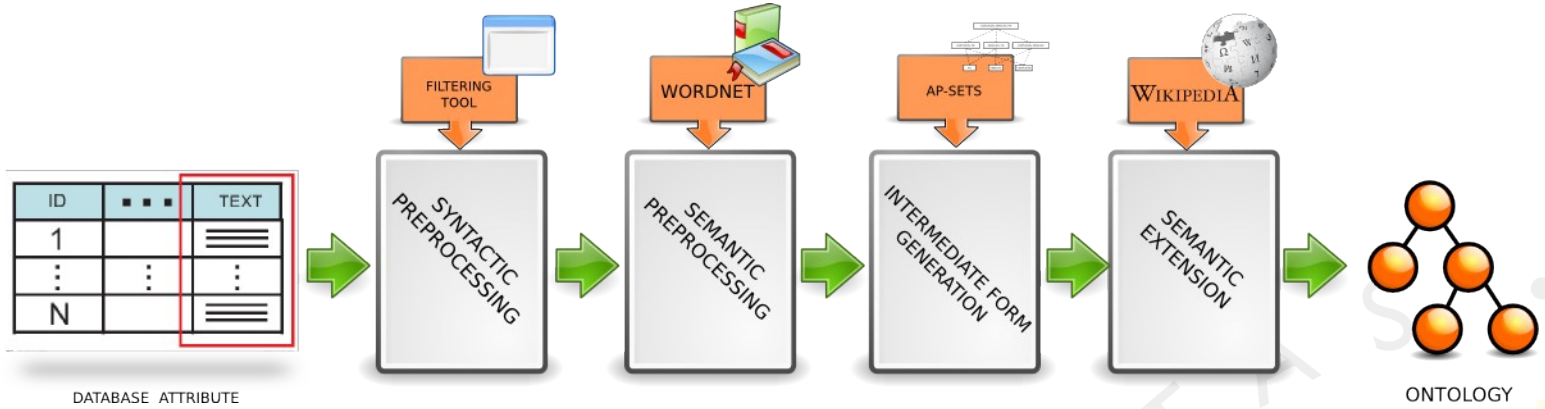
- Bolsa de palabras (BOW)
- Representación estructurada

Ejemplo: Generación de Ontologías VI



- Se seleccionan los términos apropiados de la representación intermedia
- Se obtiene información adicional de una ontología de referencia o herramienta externa.

Ejemplo: Generación de Ontologías VIII



El origen de los datos puede ser documentos de texto, o tuplas de una base de datos.

La representación intermedia va a determinar la forma de procesar los datos.

La instanciación se puede hacer con herramientas tales como WordNet o Wikipedia.

Bibliografía

• AlSumait L, Barbar D and Domeniconi C 2008 Online LDA: Adaptive topic model for mining text streams with application on topic detection and tracking. Proceedings of the IEEE International Conference on Data Mining.

• Bloehdorn, S., Blohm, S., Cimiano, P., Giesbrecht, E., Hotho, A., Lösch, U., & Völker, J. (2011). Combining Data-Driven and Semantic Approaches for Text Mining (pp. 115-142). Springer Berlin Heidelberg.

• Feldman, R., & Dagan, I. (1995, August). Knowledge Discovery in Textual Databases (KDT). In KDD (Vol. 95, pp. 112-117).

• Berry, MW. and Kogan, J. (2010). Text Mining. Applications and Theory. John Wiley & Sons

• Ejemplo LSA: http://matpalm.com/lsa_via_svd/index.html

UNIVERSIDAD DE GRANADA

decsai.ugr.es

Sesión II.4: Minería de Texto Multilingüe

DECSAI

Departamento de Ciencias de la Computación e Inteligencia Artificial

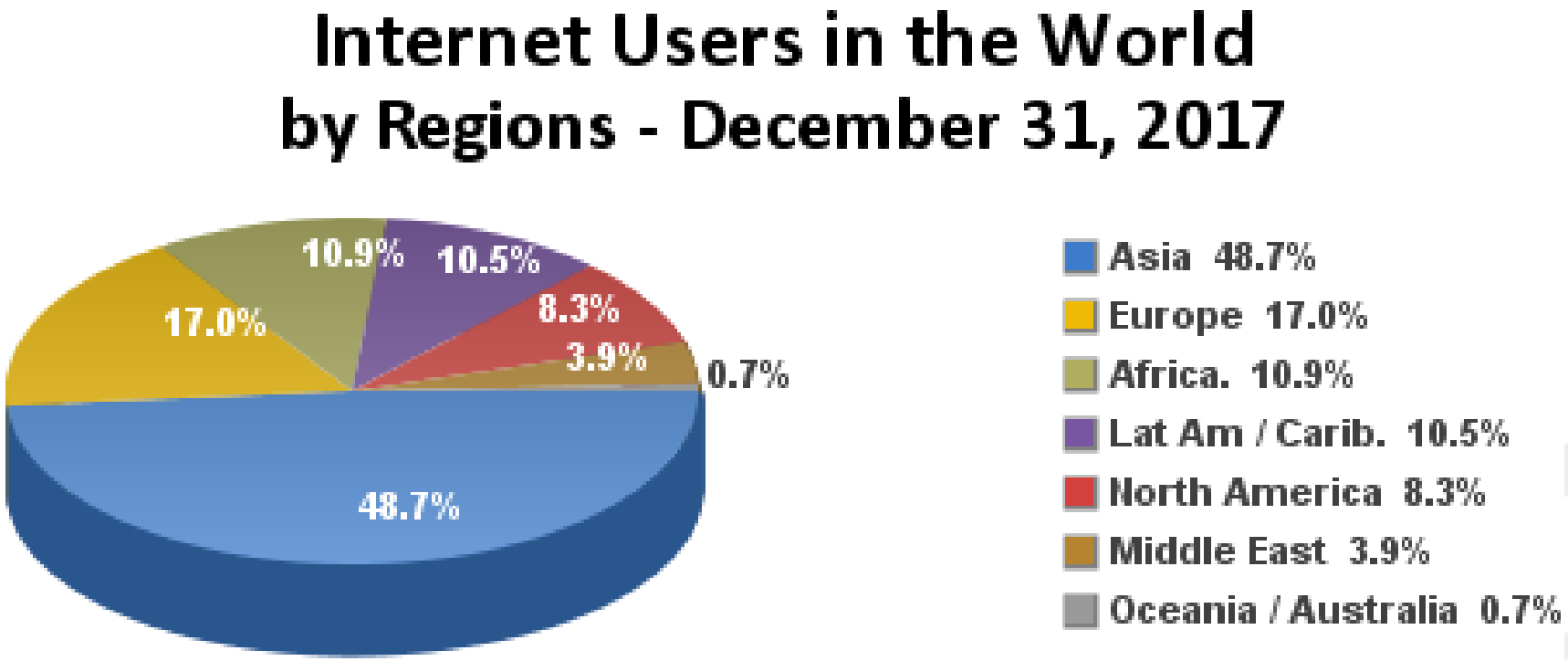
Minería de Texto Multilingüe

- Los contenidos de las páginas web se encuentran en diversos idiomas.

- En ocasiones, la información relevante para un usuario no se encuentra en su propio idioma.

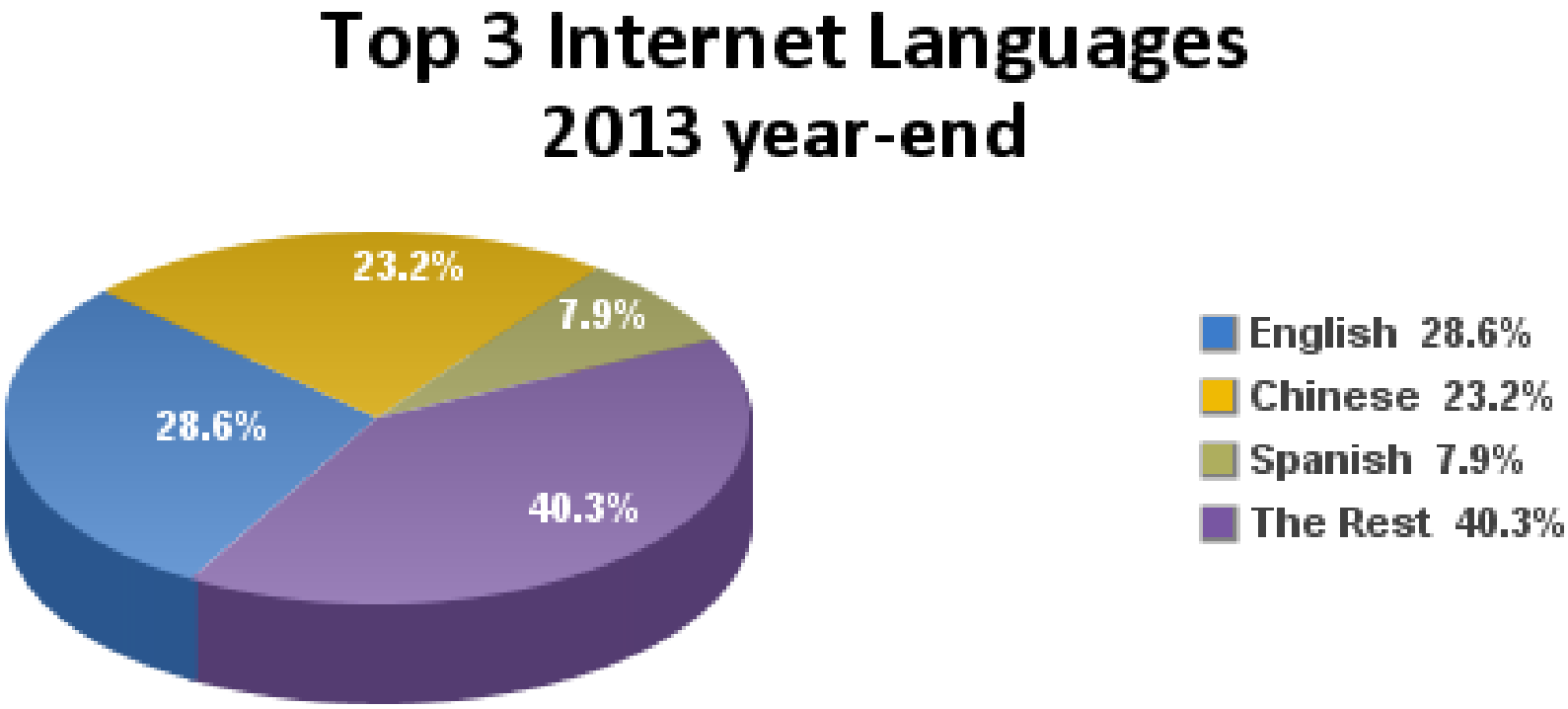
- Los Recursos Léxicos sólo existen para algunos idiomas.

Usuarios de Internet en el Mundo



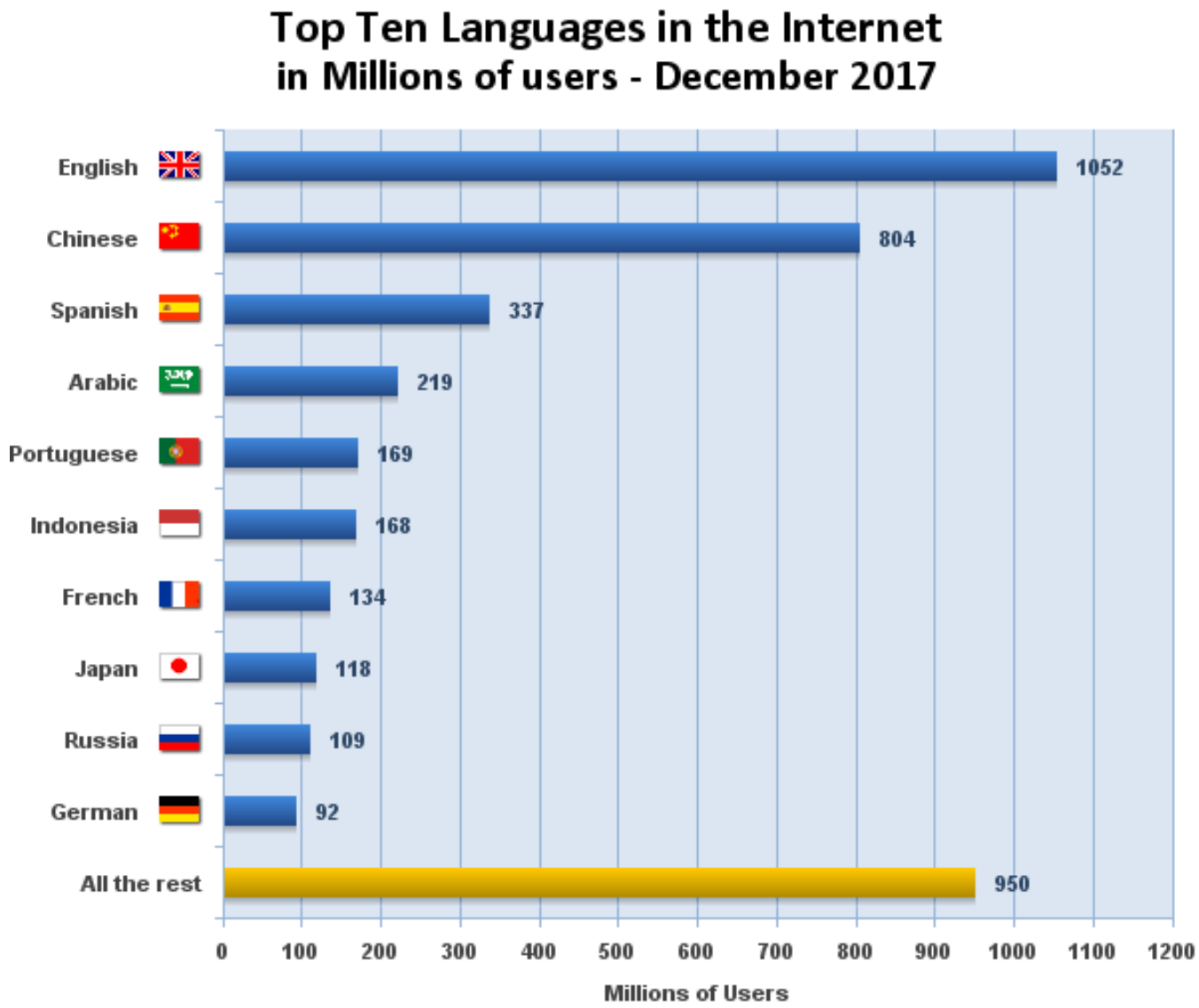
Source: Internet World Stats - www.internetworldstats.com/stats.htm
Basis: 4,156,932,140 Internet users in December 31, 2017
Copyright © 2018, Miniwatts Marketing Group

Lenguajes Usados en Internet I



Source: Internet World Stats - www.internetworldstats.com/languages.htm
Based on 2,802,478,934 estimated Internet users for December 31, 2013
Copyright © 2014, Miniwatts Marketing Group

Lenguajes Usados en Internet II



Source: Internet World Stats - www.internetworldstats.com/stats7.htm
Estimated total Internet users are 4,156,932,140 in December 31, 2017
Copyright © 2018, Miniwatts Marketing Group

Minería de Texto Multilingüe

La gran cantidad de información disponible en diversos idiomas, ha generado la necesidad de adaptar las técnicas de minería de texto a contextos multilingües.

Algunas de las aplicaciones desarrolladas son:

- Traducción automática (Machine Translation).
- Recuperación de contenido multilingüe.
- Extracción de conocimiento entre lenguajes (cross-language).
- Minería de opiniones entre lenguajes.
- Categorización de textos.
- Resumen de textos.

Cada una de estas aplicaciones presenta una serie de retos diferentes.

Resumen de Textos

El resumen de textos multilingües consiste en crear un resumen de un conjunto de documentos similares, incluyendo las frases más relevantes de dichos documentos en el resumen.

Dado un conjunto de documentos (cluster), los pasos a seguir para realizar un resumen son:

- Calcular el vector medio del cluster.
- Ordenar los documentos según su similitud al vector medio.
- Seleccionar sentencias candidatas a partir de pesos que incluyen su relevancia y posición en el documento.
- Comprobar la similitud de la sentencia con otras presentes en el resumen.
- Añadir la sentencia candidata si no es redundante.
- Añadir sentencias hasta alcanzar el tamaño de resumen deseado.

El origen multilingüe de los textos obliga a seleccionar un idioma objetivo y traducir los resultados a dicho idioma.

LSA Multilingüe

Se necesita un corpus multi-paralelo en el que los mismos textos estén traducidos a distintos idiomas (se suelen usar traducciones de la biblia y el corán, así como documentos de patentes).

Cuando se usan múltiples lenguajes, se apilan las matrices de términos-documentos. Existen diversas aproximaciones para realizar estos apilamientos:

- **Básica** [Chew and Abdelali (2007)]: Se colocan las matrices unas a continuación de otras. Las filas corresponden a términos en todos los lenguajes.

- **Tucker1** [Kolda and Bader (2009); Tucker (1966)]: Se colocan las matrices formando una tercera dimensión.

$$X_k \approx U_k S_k V^T \quad k = 1, \ldots, K.$$

- **PARAFAC2** [Harshman (1972)]: Similar a Tucker1, incluye una matriz H diagonal densa.

$$X_k \approx U_k H S_k V^T \quad k = 1, \ldots, K.$$

Minería de Texto Multilingüe

El análisis de texto multilingüe aporta el beneficio de capturar información complementaria del mismo evento a través de distintos lenguajes.

- Sobre Contenido
- Sobre Opiniones y Sentimientos

Uno de los grandes desafíos es desarrollar **herramientas**, **recursos** y **aplicaciones** de minería de texto multilingüe con el menor coste de desarrollo y tiempo posible.

Propuestas para Desarrollo de Herramientas

- Unicode.
- Teclados Virtuales.
- Modularidad.
- Clases de Token compartidas.
- Estructuras de entrada y salida uniformes.
- Simplicidad en las reglas y el lexicon.
- Compartir recursos entre lenguajes (lexica, gazetteers, grammar rules).
- Usar teoría de gramática.
- Usar Aprendizaje Automático (Machine Learning).
- No especificar en exceso.
- Minimizar el uso de herramientas específicas para un lenguaje.
- Evitar herramientas específicas para un lenguaje.

Ralf Steinberger

Recursos Multilingües I

WordNet [Fellbaum, 1998]

- http://wordnet.princeton.edu/

- Se comenzó a desarrollar en 1985 en el Laboratorio de Ciencia Cognitiva de Princeton.

- Base de Datos léxica que contiene información sobre **nombres**, **verbos**, **adjetivos** y **adverbios** en Inglés.

- Para buscar en WordNet debemos conocer el **lema** de la palabra y su **categoría gramatical** (part-of-speech/POS)

- Se organiza en **Synsets**, conjuntos de palabras sinónimas.

- Los Synsets están relacionados a través de relaciones **conceptuales**, **semánticas** y **léxicas**, estableciendo una red.

Recursos Multilingües II

WordNet [Fellbaum, 1998]

- Algunas de estas relaciones son:
  - **Hiponimia (Hyponym)**
    El término específico usado para designar un miembro de una clase. X es un hipónimo de Y, si X es algún tipo de Y.
  - **Hiperonimia (Hypernym)**
    El término general empleado para designar una clase completa de instancias específicas. Y es un hiperónimo de X, si X es algún tipo de Y.
  - **Meronimia (Meronym)**
    El nombre de un constituyente de parte de, la sustancia de, el miembro de algo. X es un merónimo de Y, si X es parte de Y.
  - **Holonimia (Holonym)**
    El nombre del todo al que hacen referencia los merónimos. Y es un holónimo de X, si X es parte de Y.
  - **Antonimia, Troponimia, Similitud, ...**

Recursos Multilingües III

EuroWordNet [Vossen, 1998]

- Base de datos que almacena WordNets en distintos idiomas.

- Proyecto Europeo para integrar WNs de 8 lenguajes europeos: ES, IT, EN, NL, FR, DE, CS, ET.

- Los distintos WNs se conectan a través de un índice Inter-Lingual-Index (ILI) que permite relacionar synsets en diferentes idiomas.

Recursos Multilingües IV

Multilingual Central Repository (MCR)

- Basado en EuroWordNet, integra en éste distintas versiones del WordNet de Princeton, y WordNets para *Castellano*, *Euskera*, *Gallego* y *Catalán*.
- Además incluye:
  - *EuroWordNet Top Concept Ontology – Enlazando* 64 conceptos definidos en la ontología con los ILI.
  - *WordNet Domains* – Extiende WN con etiquetas de dominio para cada synset, seleccionadas de un conjunto de 200 etiquetas estructuradas de forma jerárquica.
  - *Nuevas relaciones obtenidas de forma automática.*

Recursos Multilingües V

Listados de WordNets disponibles y el tipo de licencia que soportan:

Open Multilingual Wordnet

http://compling.hss.ntu.edu.sg/omw/

Wordnets in the World

http://globalwordnet.org/wordnets-in-the-world/

Recursos Multilingües VI

BabelNet [Navigli – Ponzetto, 2012]

- http://babelnet.org/
- Red semántica y ontología multilingüe lexicalizada.
- Desarrollada por Linguistic Computing Laboratory de la Universidad de la Sapienza (Roma).
- Conecta **Wikipedia** y **WordNet**, además de otros recursos como **WikiData**, **Wiktionary**, **OmegaWiki** y **Open Multilingual WordNet**.
- Proporciona conceptos y entidades lexicalizadas en diversos idiomas, así como sus relaciones semánticas.
- Está disponible una **API en Java** para acceder a los datos a través de un servicio *HTTP RESTful*.
- Cada **Babel synset** representa un sentido y contiene sinónimos en diferentes lenguajes.

Integración de Ontologías y Lexicons

Para poder establecer una comunicación y realizar un intercambio de información es necesario compartir el mismo conjunto de palabras (lexicon) y conocer el modelo subyacente a éste.

Este modelo puede representarse como una ontología, cuya función es agrupar conceptos similares, definir sus relaciones mutuas, y dar soporte a la herencia de propiedades y el razonamiento.

El lexicon y la ontología capturan diferentes tipos de información

- **Lexicon:** Información sintáctica específica del lenguaje e información morfológica.
- **Ontología:** Es independiente del lenguaje y captura el significado formal y las interrelaciones entre conceptos que no se reflejan en el lexicon.

Bibliografía

• Fellbaum, C. (1998). WordNet. Blackwell Publishing Ltd.

• Vossen, P. (1998). A multilingual database with lexical semantic networks. Kluwer Academic Publishers, Dordrecht.

• R. Navigli and S. Ponzetto. (2012). BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. Artificial Intelligence, 193, Elsevier, pp. 217-250.

• Bloehdorn, S., Blohm, S., Cimiano, P., Giesbrecht, E., Hotho, A., Lösch, U., ... & Völker, J. (2011). Combining Data-Driven and Semantic Approaches for Text Mining (pp. 115-142). Springer Berlin Heidelberg.

• Feldman, R., & Dagan, I. (1995, August). Knowledge Discovery in Textual Databases (KDT). In KDD (Vol. 95, pp. 112-117).

• Berry, MW. and Kogan, J. (2010). Text Mining. Applications and Theory. John Wiley & Sons