Big Data: fundamentos tecnológicos y aplicaciones prácticas

Cursos de Verano de la Universidad de Alicante

11–14 julio 2016

Text Mining y Web Mining

David Tomás Díaz
Depto. Lenguajes y Sistemas Informáticos
Universidad de Alicante

Universitat d'Alacant
Universidad de Alicante

Preguntas a responder hoy

- ¿Cuál es la diferencia entre Data Mining, Text Mining y Web Mining?
- ¿Cuáles son los componentes de este tipo de sistemas?
- ¿Qué aplicaciones tienen en la vida real?
- ¿Cómo podemos hacer que un texto sea entendible por un ordenador?
- ¿Qué tipo de información podemos extraer de la Web?
- Y muchas más…

Text Mining


blue
leave
moment
sing
whisper
time
size
spring
those
behind
your
picture
stare
see
through
for
may
beauty
worship
would

Contenidos

- ¿Qué es Text Mining?
- Text Mining vs Data Mining
- La variabilidad y ambigüedad del lenguaje
- La aproximación lingüística
- La aproximación estadística
- La efectividad irracional de los datos
- Representación de los documentos
- Algoritmos
- Aplicaciones

Contenidos

- **¿Qué es Text Mining?**
- Text Mining vs Data Mining
- La variabilidad y ambigüedad del lenguaje
- La aproximación lingüística
- La aproximación estadística
- La efectividad irracional de los datos
- Representación de los documentos
- Algoritmos
- Aplicaciones

¿Qué es Text Mining?

- "Minería de textos": proceso de extraer información útil de grandes conjuntos de datos textuales
- ¿Por qué es interesante?
  - Gran cantidad de datos textuales creados en diferentes redes sociales, web y otras aplicaciones centradas en la información
  - Los datos no estructurados son la forma más sencilla de datos que pueden ser creados en cualquier escenario de aplicación

Contenidos

- ¿Qué es Text Mining?
- **Text Mining vs Data Mining**
- La variabilidad y ambigüedad del lenguaje
- La aproximación lingüística
- La aproximación estadística
- La efectividad irracional de los datos
- Representación de los documentos
- Algoritmos
- Aplicaciones

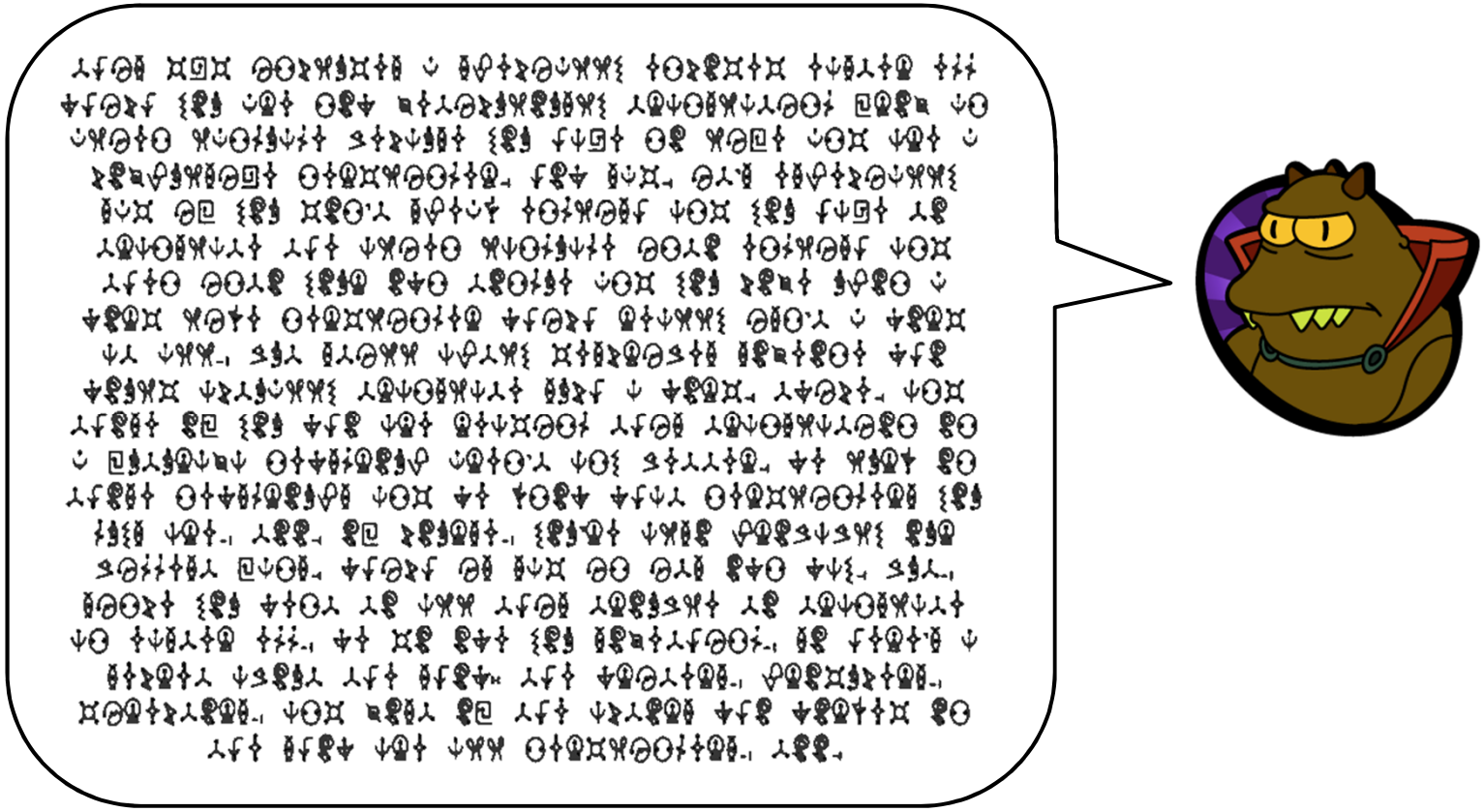
Text Mining vs Data Mining

Data Mining	Text Mining

• Busca patrones en los datos
• Información a extraer de los datos
  • Implícita (oculta)
  • Previamente desconocida
  • Difícilmente extraíble sin técnicas automáticas
	• Busca patrones en el texto
	• Información a extraer de los datos
	  • Clara y explícitamente presente en el texto
	  • No expresada de manera amigable para su procesado automático

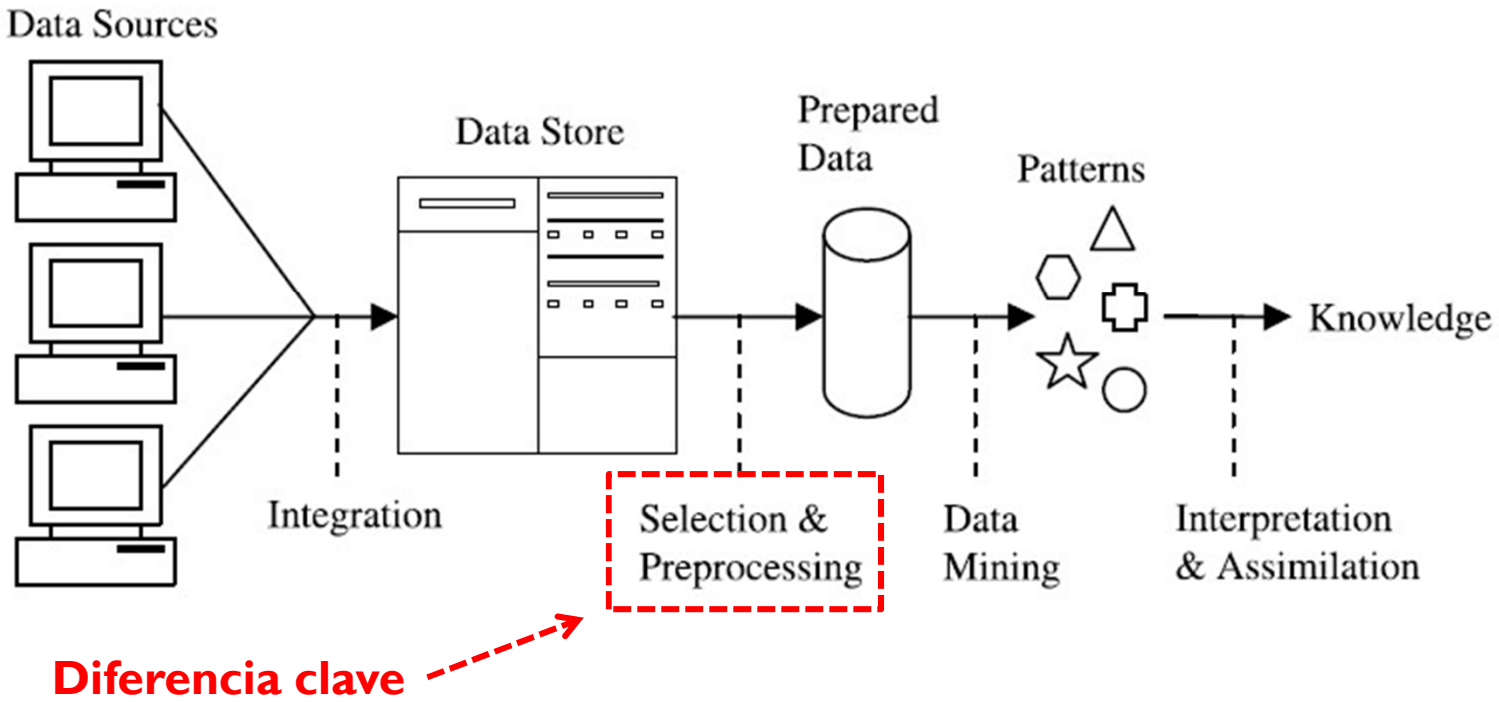
Text Mining vs Data Mining

- El texto es tan oscuro (o probablemente más) como los datos en crudo a la hora de extraer información



Text Mining vs Data Mining

- Hay una gran diferencia filosóficamente…
- … pero desde el punto de vista computacional ambos problemas son bastante similares

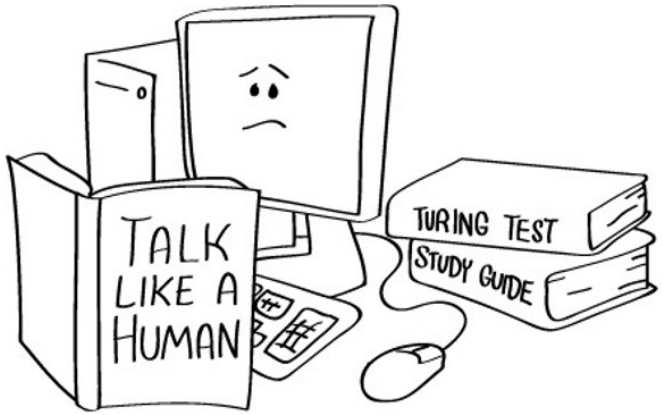


Contenidos

- ¿Qué es Text Mining?
- Text Mining vs Data Mining
- **La variabilidad y ambigüedad del lenguaje**
- La aproximación lingüística
- La aproximación estadística
- La efectividad irracional de los datos
- Representación de los documentos
- Algoritmos
- Aplicaciones

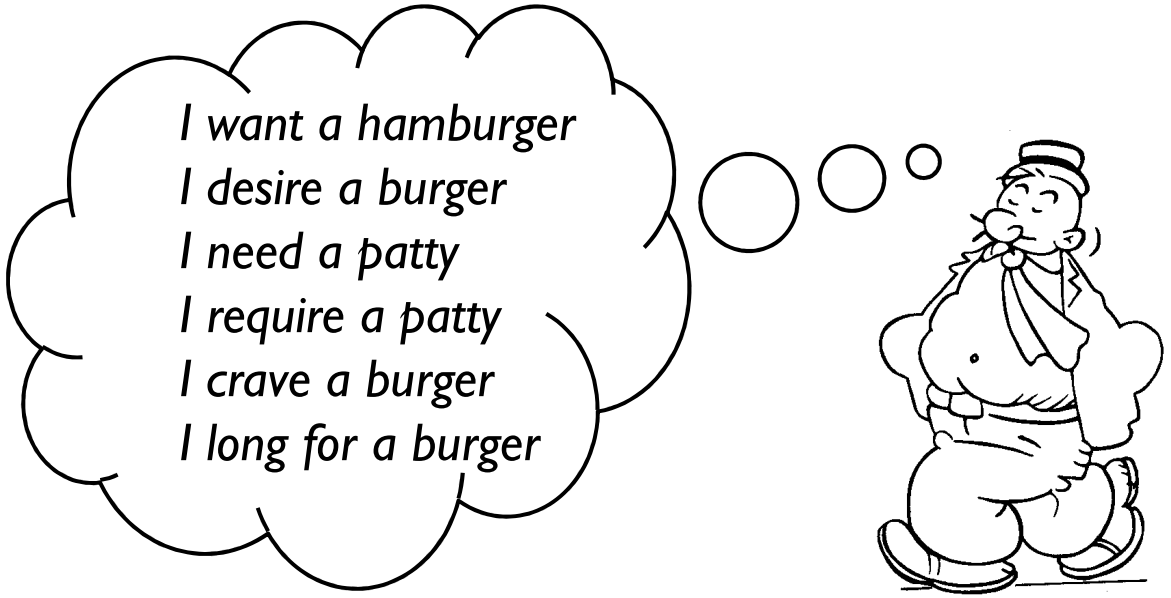
La variabilidad y ambigüedad del lenguaje

- Entender el lenguaje natural (humano) parece sencillo e intuitivo para una persona…
- … pero diversos factores afectan al rendimiento y robustez de los sistemas automáticos
- El principal problema es el fenómeno de la **sinonimia** y la **polisemia**, es decir, la variabilidad y la ambigüedad del lenguaje natural



La variabilidad y ambigüedad del lenguaje

- La variabilidad del lenguaje natural
  - También conocida como **sinonimia**
  - Formular la misma información de muchas maneras diferentes
  - Oraciones semánticamente similares pueden ser completamente diferentes desde un punto de vista léxico



La variabilidad y ambigüedad del lenguaje

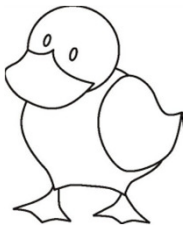
- La ambigüedad del lenguaje natural
  - También conocida como **polisemia**
  - Algo es ambiguo cuando puede ser entendido de dos o más maneras o sentidos diferentes

La variabilidad y ambigüedad del lenguaje

- La ambigüedad del lenguaje natural
  - También conocida como **polisemia**
  - Algo es ambiguo cuando puede ser entendido de dos o más maneras o sentidos diferentes

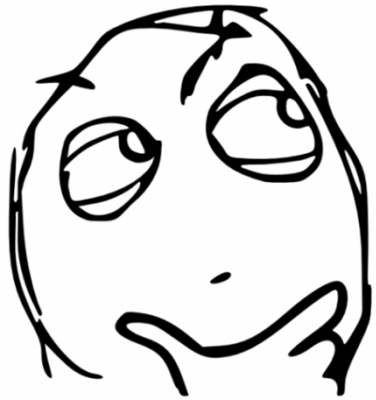
La frase *I made her duck* tiene al menos 5 significados

- Cociné un pato para ella
- Cociné un pato que era de ella
- Fabriqué el pato (¿de escayola?) que tiene ella
- Hice que agachara rápidamente la cabeza
- Agité mi varita mágica y la transformé en un pato



La variabilidad y ambigüedad del lenguaje

¿Cómo podemos manejar estos problemas y representar el lenguaje de manera tratable para un ordenador?



La variabilidad y ambigüedad del lenguaje

Aproximación lingüística
vs
Aproximación estadística

Contenidos

- ¿Qué es Text Mining?
- Text Mining vs Data Mining
- La variabilidad y ambigüedad del lenguaje
- **La aproximación lingüística**
- La aproximación estadística
- La efectividad irracional de los datos
- Representación de los documentos
- Algoritmos
- Aplicaciones

La aproximación lingüística

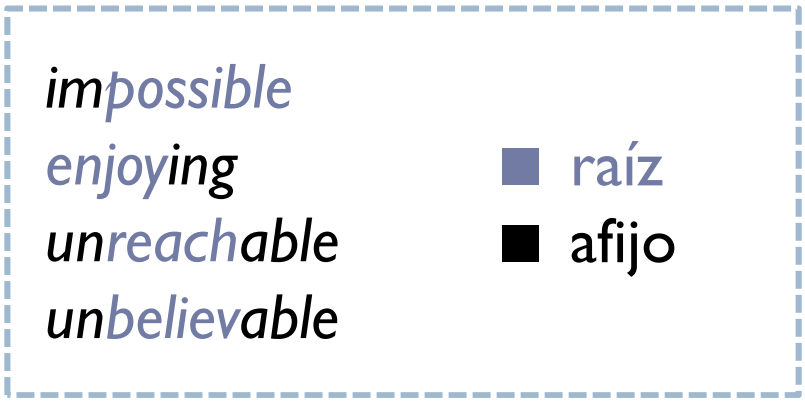
- También conocida como **simbólica**
- Almacenar de manera explícita los hechos/conocimiento
- Análisis de diferentes niveles lingüísticos
  - Fonología
  - Morfología
  - Sintaxis
  - Semántica
  - Pragmática

La aproximación lingüística

- También conocida como **simbólica**
- Almacenar de manera explícita los hechos/conocimiento
- Análisis de diferentes niveles lingüísticos
  - Fonología
  - **Morfología**
  - **Sintaxis**
  - **Semántica**
  - Pragmática

La aproximación lingüística

- Morfología
  - El estudio de cómo las palabras se componen de morfemas
  - Dos grandes clases de morfemas
    - **Raíz**: morfema principal de la palabra, proporciona significado
    - **Afijo**: piezas que se combinan con la raíz para modificar el significado y las funciones gramaticales



La aproximación lingüística

- Morfología
  - Etiquetado morfológico (*Part-of-Speech tagging*)
    - Identificar la función gramatical particular de una palabra en un texto (nombre, verbo, adjetivo, …)

The grand jury commented on a number of other topics.

```
The the DT 1
grand grand JJ 0.832524
jury jury NN 1
commented comment VBD 0.954545
on on IN 0.971769
a 1 Z 0.99998
number number NN 0.998704
of of IN 0.999898
other other JJ 0.632399
topics topic NNS 1
. . Fp 1
```

La aproximación lingüística

- Morfología
  - Etiquetado morfológico (*Part-of-Speech tagging*)
    - Ejemplo de uso de Freeling
      - http://nlp.lsi.upc.edu/freeling/demo/demo.php
      - *Select output > PoS Tagging*

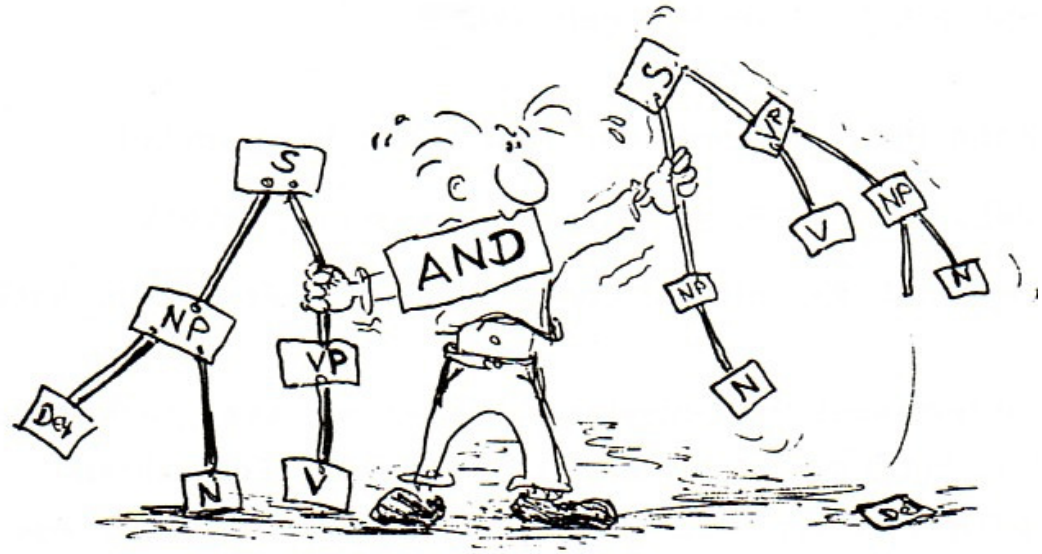
> Hoy es jueves, 14 de julio de 2016. Son las 10 de la mañana, más o menos. Estoy en un curso de Big Data en la Universidad de Alicante, en el campus de San Vicente del Raspeig. Estamos a cuarenta grados. David le pone ganas, pero yo estoy pensando en irme a la playa y tomarme tres cervezas.

La aproximación lingüística

- Morfología
  - Aplicaciones
    - Es un primer paso en muchas tareas prácticas de Text Mining
    - El proceso de análisis sintáctico necesita saber si una palabra es un nombre o un verbo antes de llevarse a cabo
    - Encontrar nombres y sus relaciones para los sistemas de Extracción de Información
    - Identificar raíces y lemas para los sistemas de Recuperación de Información
    - Eliminar potenciales palabras vacías (*stopwords*) para la reducción de dimensionalidad
    - …

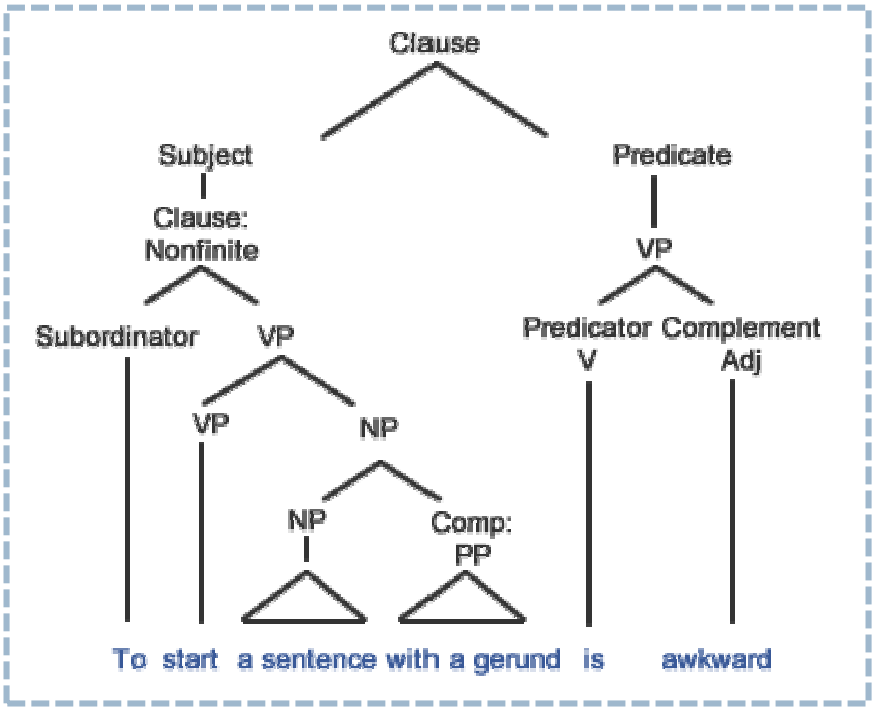
La aproximación lingüística

- Sintaxis
  - El análisis sintáctico se centra en la construcción de oraciones
  - La estructura sintáctica indica cómo las palabras se relacionan unas con otras



La aproximación lingüística

- Sintaxis
  - Análisis completo
    - Obtiene la estructura anidada de la frase
    - Proporciona el papel de los constituyentes en la oración principal



La aproximación lingüística

- Sintaxis
  - Análisis superficial
    - También conocido como *chunking*
    - Análisis de una oración que identifica los constituyentes (sintagmas nominales, verbos, sintagmas verbales,...)
    - Divide el texto plano en secuencias de palabras semánticamente relacionadas
    - No especifica su estructura interna ni su papel en la oración principal

```
[NP Jack and Jill] [VP went] [ADVP up] [NP the hill]
    [VP to fetch] [NP a pail] [PP of] [NP water]
```

La aproximación lingüística

- Sintaxis
  - Análisis superficial
    - La tarea es más sencilla de resolver por un ordenador (y usada más frecuentemente) que el análisis completo

Análisis completo

```
(S (NP He)
  (VP reckons
    (S (NP the current account deficit)
      (VP (VP will narrow)
        (PP to (NP only 1.8 billion))
        (PP in (September))))))
```

Análisis superficial

```
[NP He] [VP reckons] [NP the current account deficit] [VP will narrow] [PP to] [NP only 1.8 billion] [PP in] [NP September]
```

La aproximación lingüística

- Sintaxis
  - Análisis superficial
    - Ejemplo de uso de Freeling
      - http://nlp.lsi.upc.edu/freeling/demo/demo.php
      - *Select output > Shallow Parsing*

> Hoy es jueves, 14 de julio de 2016. Son las 10 de la mañana, más o menos. Estoy en un curso de Big Data en la Universidad de Alicante, en el campus de San Vicente del Raspeig. Estamos a cuarenta grados. David le pone ganas, pero yo estoy pensando en irme a la playa y tomarme tres cervezas.

La aproximación lingüística

- Sintaxis
  - Aplicaciones
    - Preparar el texto para la interpretación semántica
    - Búsqueda de respuestas (*Question Answering*)
    - Extracción de información (*Information Extraction*)
    - Generación del lenguaje (*Language Generation*)
    - Traducción automática (*Machine Translation*)
    - …

La aproximación lingüística

- Semántica
  - Estudia el significado de las expresiones lingüísticas
  - Temas que estudia la semántica
    - Significado léxico
      - El significado de las palabras individuales
    - Principio de composicionalidad
      - El significado de una unidad mayor se obtiene a partir del significado de sus partes
    - Resolución de la ambigüedad
      - Cómo determinar en un contexto el sentido de una expresión lingüística que tiene diferentes significados
      - Desambiguación del sentido de las palabras (*Word Sense Disambiguation*)

La aproximación lingüística

- Semántica
  - Desambiguación del sentido de las palabras
    - Ejemplo de uso de Freeling
      - http://nlp.lsi.upc.edu/freeling/demo/demo.php
      - *Analysis options > WN sense annotation*

> Hoy es jueves, 14 de julio de 2016. Son las 10 de la mañana, más o menos. Estoy en un curso de Big Data en la Universidad de Alicante, en el campus de San Vicente del Raspeig. Estamos a cuarenta grados. David le pone ganas, pero yo estoy pensando en irme a la playa y tomarme tres cervezas.

La aproximación lingüística

- Semántica
  - WordNet
    - La base de datos léxica más usada para el idioma inglés
    - Ayuda para resolver el problema de la sinonimia
    - También disponible en otros idiomas, incluyendo el castellano (*EuroWordNet*)
    - Los ítems léxicos están categorizados en (más de 115.000) *synsets*
    - Un *synset* consta de un conjunto de sinónimos, una definición de diccionario (glosa) y algunos ejemplos de uso
    - Las relaciones semánticas se representan como redes que las aplicaciones pueden recorrer para encontrar sinónimos, antónimos, hiperónimos, hipónimos, …

La aproximación lingüística

- Semántica
  - WordNet

Noun

- S: (n) **table**, tabular array **(a set of data arranged in rows and columns)** ***"see table 1"***
  - direct hyponym / full hyponym
  - member meronym
  - direct hypernym / inherited hypernym / sister term
  - derivationally related form
- S: (n) **table** (a piece of furniture having a smooth flat top that is usually supported by one or more vertical legs) *"it was a sturdy table"*
- S: (n) **table** (a piece of furniture with tableware for a meal laid out on it) *"I reserved a table at my favorite restaurant"*
- S: (n) mesa, **table** (flat tableland with steep edges) *"the tribe was relatively safe on the mesa but they had to descend into the valley for water"*
- S: (n) **table** (a company of people assembled at a table for a meal or game) *"he entertained the whole table with his witty remarks"*
- S: (n) board, **table** (food or meals in general) *"she sets a fine table"; "room and board"*

Verb

- S: (v) postpone, prorogue, hold over, put over, **table**, shelve, set back, defer, remit, put off (hold back to a later time) *"let's postpone the exam"*
- S: (v) **table**, tabularize, tabularise, tabulate (arrange or enter in tabular form)

La aproximación lingüística

- Semántica
  - WordNet
    - Ejemplo de uso de WordNet
      - http://wordnetweb.princeton.edu/perl/webwn

- dog
- pick up
- Kennedy
- Spain

La aproximación lingüística

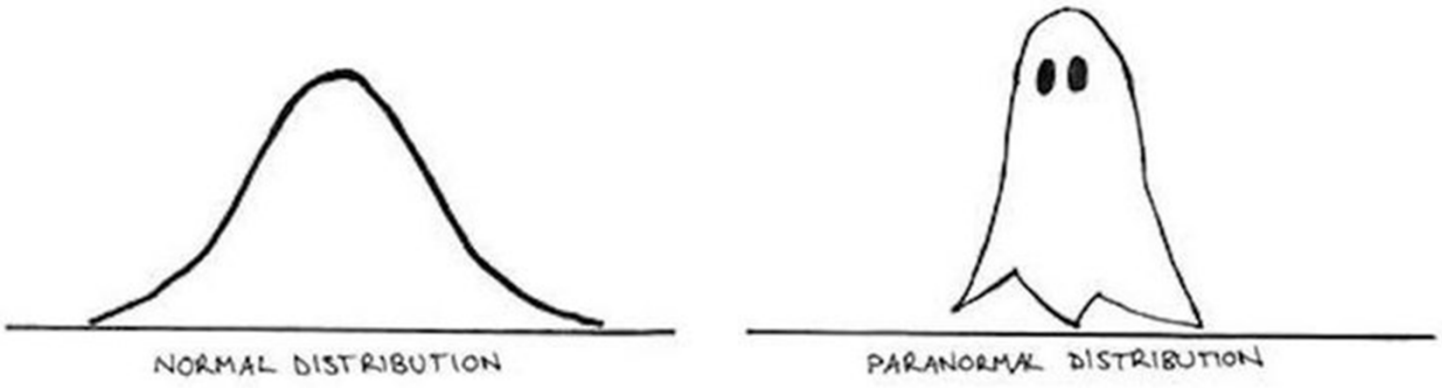
- Semántica
  - Aplicaciones
    - Identificación de sinónimos
    - Desambiguación del sentido de las palabras
    - Búsqueda de respuestas (*Question Answering*)
    - Traducción automática (*Machine Translation*)
    - Etiquetado de roles semánticos (*Semantic Role Labelling*)
    - Poblar ontologías
    - ...

Contenidos

- ¿Qué es Text Mining?
- Text Mining vs Data Mining
- La variabilidad y ambigüedad del lenguaje
- La aproximación lingüística
- **La aproximación estadística**
- La efectividad irracional de los datos
- Representación de los documentos
- Algoritmos
- Aplicaciones

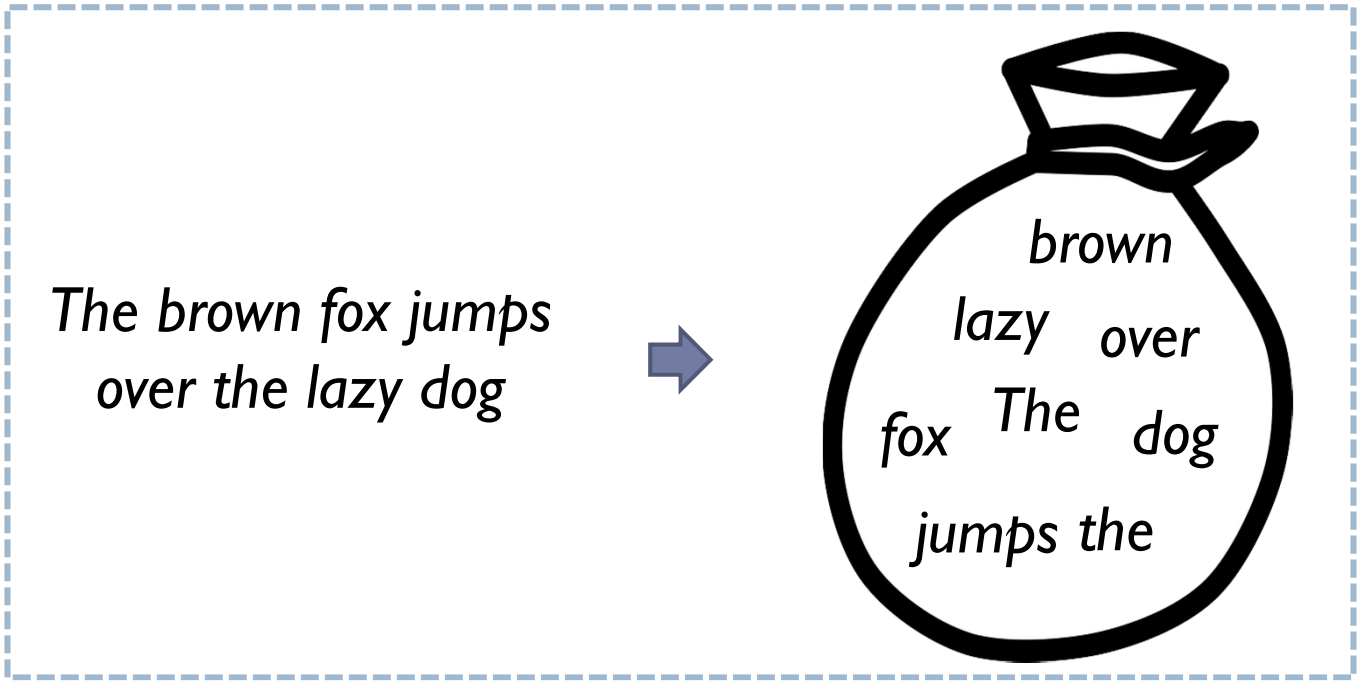
La aproximación estadística

- Usa modelos estadísticos que expresan la probabilidad de una observación particular basados en grades conjuntos de datos textuales (corpus)
- No tiene en cuenta el orden, estructura o significado
- Captura regularidades en las expresiones lingüísticas
- Trata la entrada como si fueran datos, no lenguaje



La aproximación estadística

- Formas de representar el texto
  - Bolsa de palabras (*Bag-of-words*)
    - Un documento se considera un conjunto de palabras, sin tener en cuenta su orden o estructura gramatical



La aproximación estadística

- Formas de representar el texto
  - Bi-gramas, tri-gramas, … *n*-gramas
    - Extraer todas las secuencias de dos, tres… *n* palabras en el texto

The brown fox jumps over the lazy dog

Bi-grams

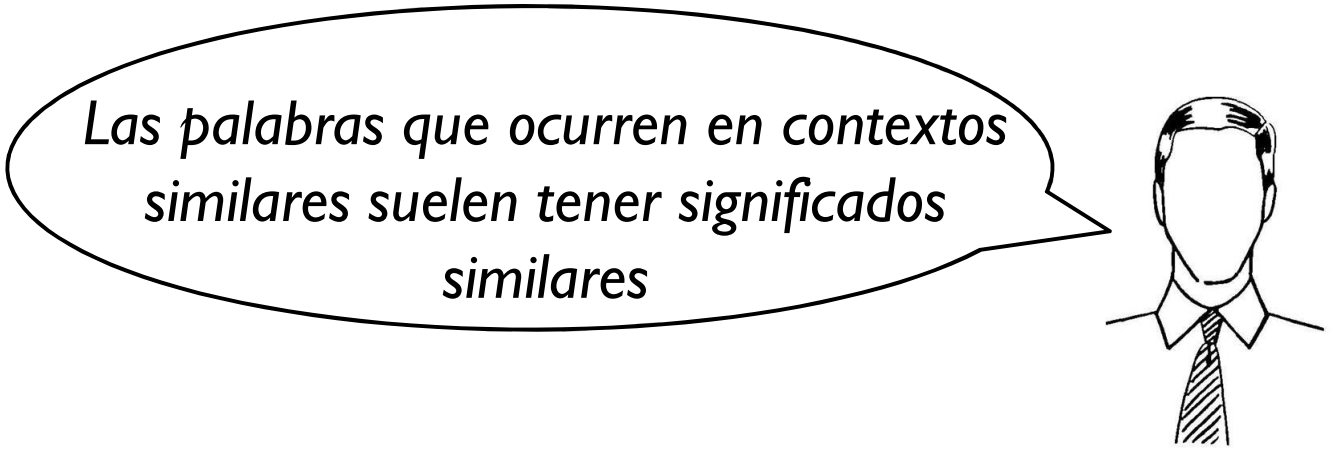
- *The brown*
- *brown fox*
- *fox jumps*
- *jumps over*
- *over the*
- *the lazy*
- *lazy dog*

Tri-grams

- *The brown fox*
- *brown fox jumps*
- *fox jumps over*
- *jumps over the*
- *over the lazy*
- *the lazy dog*

La aproximación estadística

- Semántica distribucional
  - Que no haya análisis lingüístico no implica que no haya información semántica (hipótesis distribucional)



Zellig Harris (1954)

La aproximación estadística

- Semántica distribucional
  - Que no haya análisis lingüístico no implica que no haya información semántica (hipótesis distribucional)

¿Qué significa ***fluzar*** y ***ogigi***?

María había querido la escritura de la casa desde hacía años.
Después de que Jesús finalmente le ***fluzara*** *la casa a María, ella le* ***fluzó*** *a Isabel su dúplex.*

Hay una botella de ***ogigi*** *en la mesa*
A todo el mundo le gusta el ***ogigi****.*
El ***ogigi*** *te emborracha.*
El ***ogigi*** *está hecho de maíz.*

La aproximación estadística

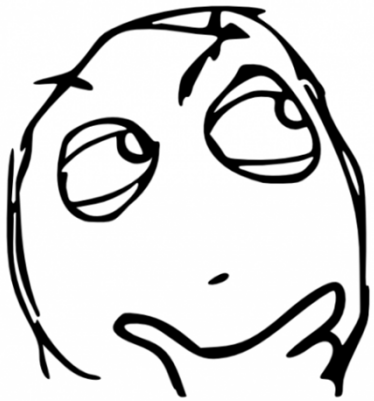
- Semántica distribucional
  - Word2vec
    - http://rare-technologies.com/word2vec-tutorial/
  - TextRazor
    - https://www.textrazor.com/
  - Cognitive Computation Group
    - http://cogcomp.cs.illinois.edu/page/demos/

Contenidos

- ¿Qué es Text Mining?
- Text Mining vs Data Mining
- La variabilidad y ambigüedad del lenguaje
- La aproximación lingüística
- La aproximación estadística
- **La efectividad irracional de los datos**
- Representación de los documentos
- Algoritmos
- Aplicaciones

La efectividad irracional de los datos

¿Puede la habilidad para procesar grandes cantidades de texto compensar por el uso de técnicas de análisis relativamente sencillas?



La efectividad irracional de los datos

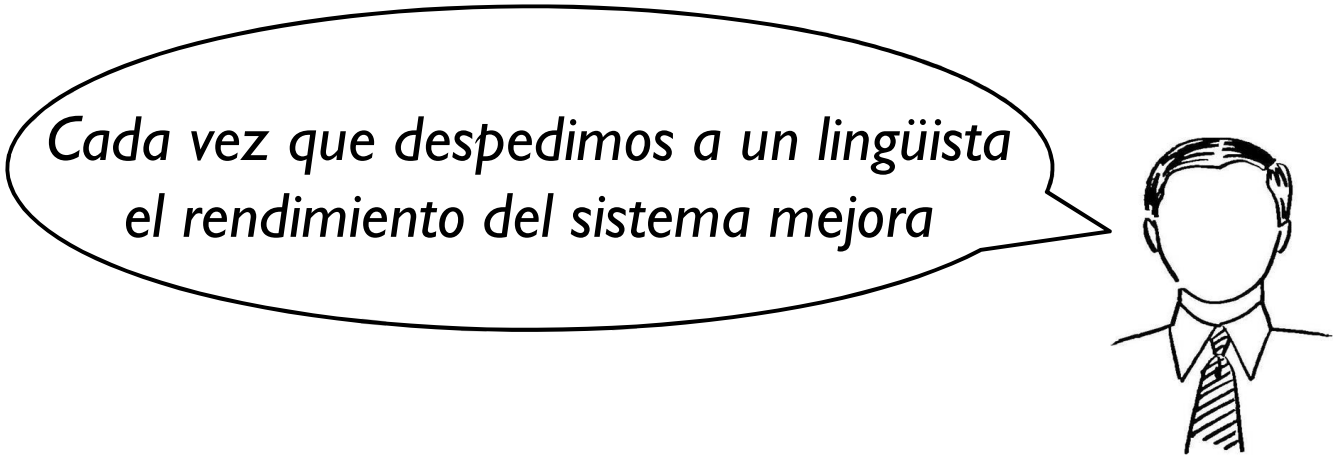
Para muchas tareas, las palabras y la combinación de éstas proporcionan toda la representación necesaria para aprender del texto

La efectividad irracional de los datos

- Problemas de la aproximación lingüística
  - Pueden haber cientos de miles de palabras diferentes y una variedad grande de construcciones gramaticales
  - Cada día surgen nuevas palabras y sus usos antiguos se van modificando
  - No podemos reducir lo que queremos decir a la combinación libre de unas pocas primitivas abstractas
  - La inferencia sobre modelos sofisticados y la extracción de características lingüísticas complejas son costosas
  - Diferentes idiomas requieren diferentes herramientas…
  - … y también los lenguajes informales requieren diferentes herramientas…

La efectividad irracional de los datos

- Problemas de la aproximación lingüística



Frederick Jelinek (1988)

La efectividad irracional de los datos

- Beneficios de la aproximación estadística
  - Hay evidencias crecientes, al menos en el procesamiento de texto, de que la cantidad de datos importa más que los atributos y algoritmos seleccionados
  - Tiene sentido aprovecharse de la cantidad de datos que nos rodean (¡Big Data!)
  - El uso de características superficiales a nivel de palabra, junto con modelos simples, superan en la mayoría de los casos a modelos más sofisticados con características más complejas pero un menor número de datos
  - La aproximación estadística es independiente del dominio y (mayoritariamente) del idioma

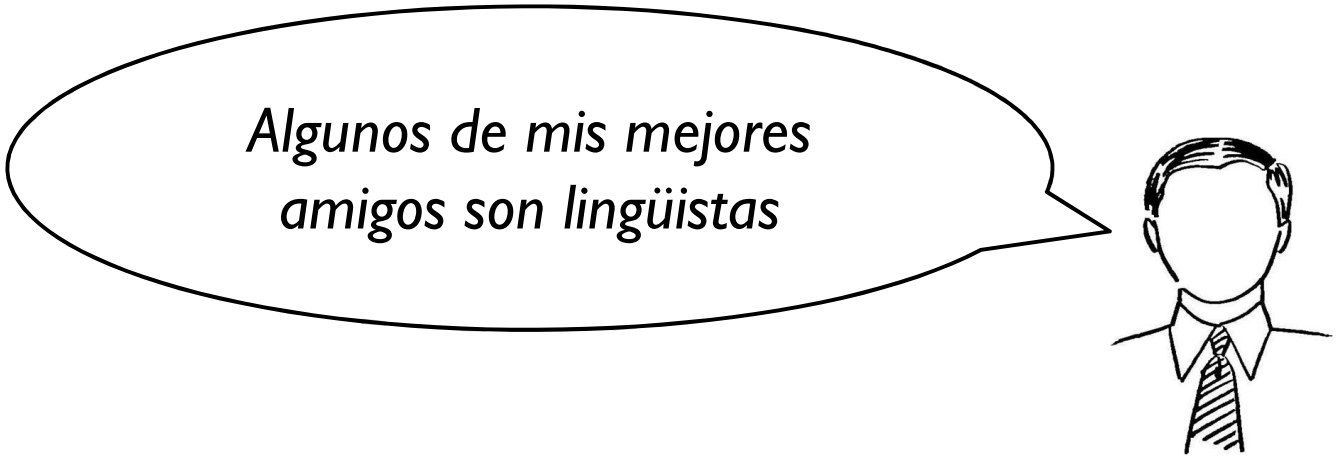
La efectividad irracional de los datos

- No obstante…
  - La aproximación lingüística juega un papel importante a la hora de obtener una representación semántica del texto más rica
    - En dominios específicos (Ej. biomedicina)
    - En tareas específicas (Ej. extracción de conocimiento de la Web)

- "The Parable of Google Flu: Traps in Big Data Analysis", Lazer et al. (2014)
- IBM Watson

La efectividad irracional de los datos

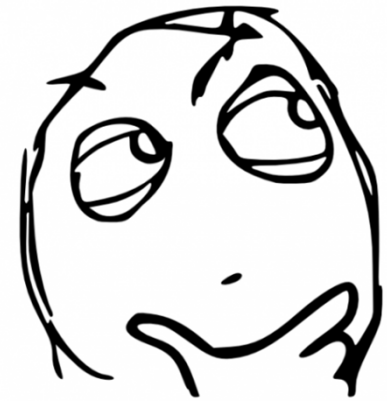
- No obstante…



Frederick Jelinek (2004)

La efectividad irracional de los datos

¿Qué deberíamos hacer entonces?



La efectividad irracional de los datos

Abrazar ambos puntos de vista

- Procesamiento superficial (estadístico) sobre textos no restringidos en dominio abierto
- Procesamiento profundo (lingüístico) sobre textos en dominios específicos

Contenidos

- ¿Qué es Text Mining?
- Text Mining vs Data Mining
- La variabilidad y ambigüedad del lenguaje
- La aproximación lingüística
- La aproximación estadística
- La efectividad irracional de los datos
- **Representación de los documentos**
- Algoritmos
- Aplicaciones

Representación de los documentos

- Un recordatorio del proceso de Data Mining
  - Hay un conjunto inicial de datos (instancias)
  - Estas instancias deben estar en un formato apropiado para ser procesado por un algoritmo de Data Mining
    - Cada instancia se representa mediante un vector de características que representan las propiedades de la instancia

Características

Instancias

Id	Reembolso	Estado Civil	Salario	Fraude
1	Sí	Soltero	125.000€	No
2	No	Casado	100.000€	No
3	No	Soltero	70.000€	No
…	…	…	…	…

Representación de los documentos

- Representación mediante bolsa de palabras
  - Es la forma más habitual de representar el texto
  - Simple y útil para muchas tareas de Text Mining
  - Cualquier documento *D* se representa como una lista de términos *t* y sus pesos asociados *p*

$$D = [(t_1;p_1),(t_2;p_2),\ldots,(t_n;p_n)]$$

t_i = término

p_i = medida de la importancia de un término a la hora de representar la información contenida en el documento

Representación de los documentos

- Representación mediante bolsa de palabras
  - Podríamos usar muchos otros tipos de atributos para representar el texto
    - Bi-gramas
    - Tri-gramas
    - Etiquetado gramatical (*POS-tags*)
    - Sintagmas nominales
    - Términos multipalabra
    - *Synsets*
    - Entidades nombradas
    - …

Representación de los documentos

Como experto en Text Mining debes decidir los atributos más útiles para cada tarea

Representación de los documentos

- Cómo transformar cada instancia en un vector de características
  - Obtención de palabras
  - Normalización / Pre-proceso
  - Esquema de pesado

Representación de los documentos

- Obtención de palabras
  - Identificar palabras individuales (*tokens*)

The Netherlands earned sweet revenge on Spain on Friday at the Fonte Nova in Salvador, hammering Spain 5-1 to put an emphatic coda on their loss in the 2010 World Cup finals.

The
Netherlands
earned
sweet
…

Representación de los documentos

- Normalización / Pre-proceso
  - Convertir las palabras a sus formas normalizadas
    - Paso a minúsculas

The → *the* ; *NASA*→ *nasa*; *Claude Shannon* → *claude shannon*

    - Obtención de lemas (*lemma*)

jumps → *jump* ; *jumping* → *jump*; *jumped* → *jump*

    - Obtención de raíces (*stem*)

computer → *comput* ; *computation* → *comput*; *compute* → *comput*

Representación de los documentos

- Normalización / Pre-proceso
  - Eliminación de palabras vacías (*stopwords*)
    - Eliminar términos comunes

> *The Netherlands earned sweet revenge on Spain on Friday at the Fonte Nova in Salvador, hammering Spain 5-1 to put an emphatic coda on their loss in the 2010 World Cup finals.*

    - Es habitual el uso de la lista de 571 palabras del sistema SMART

> *a, a's, able, about, above, according, accordingly, across, actually, after, afterwards, again, against, ain't, all, allow, allows, almost, alone, along, already, also, although, always, am, among…*

Representación de los documentos

- Normalización / Pre-proceso
  - Selección de características
    - Mantener sólo las palabras más representativas del vocabulario
    - Eliminar ruido y anomalías
    - Hay algoritmos que no pueden trabajar con un número elevado de características
    - Diferentes aproximaciones (automáticas)
      - Frecuencia del término
      - Frecuencia del documento
      - Ganancia de información (*Information Gain*)
      - Información mutua (*Mutual Information*)
      - Chi-square (χ^2)
      - …

Representación de los documentos

- Esquema de pesado
  - Medida de la importancia de un término a la hora de representar la información contenida en el documento
  - Cada término en el vector de características recibe un valor numérico
  - Existen numerosos esquemas de pesado
    - Ocurrencia del término
    - Frecuencia del término (*TF*)
    - Inversa de la frecuencia del documento (*IDF*)
    - TF-IDF
    - …

Representación de los documentos

- Esquema de pesado
  - Ocurrencia del término
    - Asignación binaria
    - Los términos ocurren (1) o no ocurren (0) en el documento

Término	Peso
best	1
dog	0
is	1
one	1
over	0
the	1
this	1

Representación de los documentos

- Esquema de pesado
  - Frecuencia del término (*Term Frequency*)
    - Las palabras repetidas están más relacionadas con el contenido
    - TF_i = número de veces que el término t_i aparece en el documento

Término	Peso
best	1
dog	0
is	1
one	2
over	0
the	1
this	1

Representación de los documentos

- Esquema de pesado
  - Inversa de la frecuencia del documento (*Inverse Document Frequency*)
    - Los términos poco comunes son más importantes
    - IDF_i = la inversa del número de documentos que contienen el término t_i con respecto al total de documentos existentes

El término *best* aparece en 3 de los10 documentos del corpus

$$IDF_{best} = 10/3 = 3{,}33$$

El término *one* aparece en 8 de los 10 documentos del corpus

$$IDF_{one} = 10/8 = 1{,}25$$

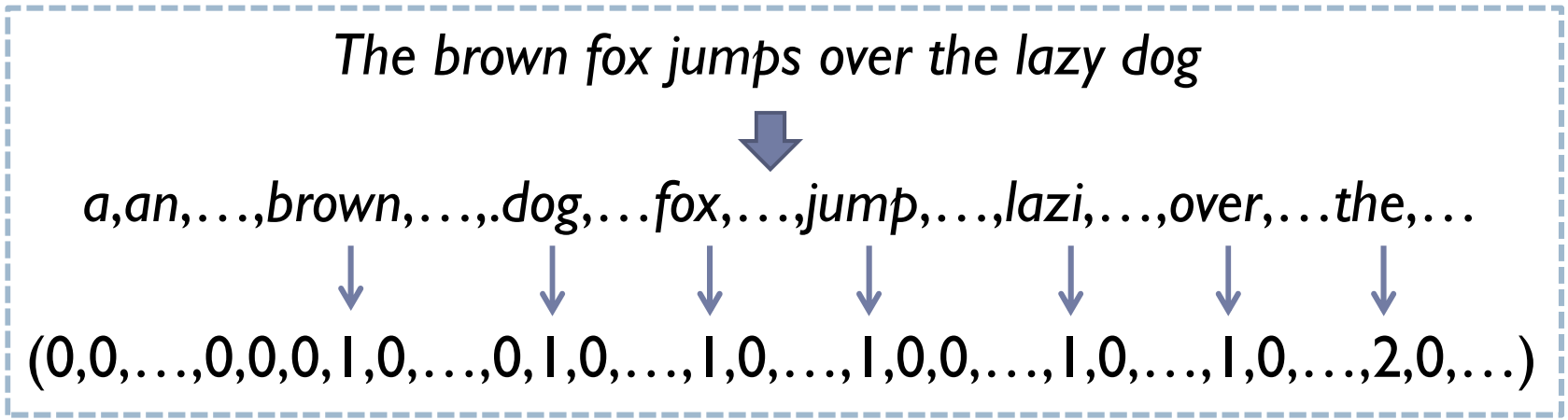
Representación de los documentos

- Esquema de pesado
  - TF-IDF
    - Más peso a palabras frecuentes que aparecen en pocos documentos
    - Combinación de *TF* e *IDF (TF-IDF = TF · IDF)*

Término	TF	IDF	TF-IDF
best	1	3,33	3,33
dog	0	5	0
is	1	1,11	1,11
one	2	1,25	2,5
over	0	10	0
the	1	1	1
this	1	1,43	1,43

Representación de los documentos

- Resumiendo…
  1. Obtener palabras y pre-procesar/normalizar
  2. Enumerar todos los términos en el corpus completo
  3. Eliminar duplicados y ordenar los términos
  4. Convertir cada término en su valor (esquema de pesado)
  5. Crear un vector de características cuyo i-ésimo valor es el peso del i-ésimo término



Representación de los documentos

- Propiedades de los vectores de características generados
  - El número de características (palabras) en un conjunto de datos puede ser muy amplio (cientos de miles)…
  - … pero un documento concreto puede contener sólo unos pocos cientos de ellas
  - Los datos textuales son dispersos y de alta dimensionalidad
    - Un gran número de características, la mayoría de ellas ocurriendo muy pocas veces
    - La mayoría de valores son 0
    - Alta proporción de características ruidosas e irrelevantes
  - La selección de características es un elemento importante al trabajar con textos

Contenidos

- ¿Qué es Text Mining?
- Text Mining vs Data Mining
- La variabilidad y ambigüedad del lenguaje
- La aproximación lingüística
- La aproximación estadística
- La efectividad irracional de los datos
- Representación de los documentos
- **Algoritmos**
- Aplicaciones

Algoritmos

- ¡Básicamente los mismos que usamos para Data Mining!
  - Supervisados
    - Clasificación
    - Regresión
  - No supervisados
    - Clustering
    - Reglas de asociación

Algoritmos

- ¡Básicamente los mismos que usamos para Data Mining!
  - Naïve Bayes
  - Árboles de decisión
  - Redes neuronales
  - Razonamiento basado en ejemplos
  - Separadores lineales
  - …

Contenidos

- ¿Qué es Text Mining?
- Text Mining vs Data Mining
- La variabilidad y ambigüedad del lenguaje
- La aproximación lingüística
- La aproximación estadística
- La efectividad irracional de los datos
- Representación de los documentos
- Algoritmos
- **Aplicaciones**

Aplicaciones

- ¿Qué podemos hacer con Text Mining?
  - Sistemas de recomendación (*Recommender systems*)
  - Minería de opiniones (*Opinion mining*)
  - Clasificación de textos (*Text Categorisation*)
  - Recuperación de información (*Information Retrieval*)
  - Modelado de temas (*Topic Modelling*)
  - Extracción de información (*Information Extraction*)
  - Resúmenes automáticos (*Text Summarisation*)
  - Atribución de autoría (*Authority Attribution*)
  - … y mucho más…

Aplicaciones

- ¿Qué podemos hacer con Text Mining?
  - Sistemas de recomendación (*Recommender systems*)
  - Minería de opiniones (*Opinion mining*)
  - **Clasificación de textos (*Text Categorisation*)**
  - Recuperación de información (*Information Retrieval*)
  - Modelado de temas (*Topic Modelling*)
  - Extracción de información (*Information Extraction*)
  - Resúmenes automáticos (*Text Summarisation*)
  - Atribución de autoría (*Authority Attribution*)
  - … y mucho más…

Clasificación de textos

- Clasificar documentos en un conjunto predefinido de clases
  - Clasificar noticias como `deportes`, `sucesos` o `política`
  - Clasificar nombres de compañías por su área de negocio
  - Clasificar correo electrónico como `spam`
  - Clasificar correos al personal técnico como `mac`, `linux`, `windows` u `otro`
  - Clasificar críticas cinematográficas como `buena`, `mala` o `neutral`
  - …

Clasificación de textos

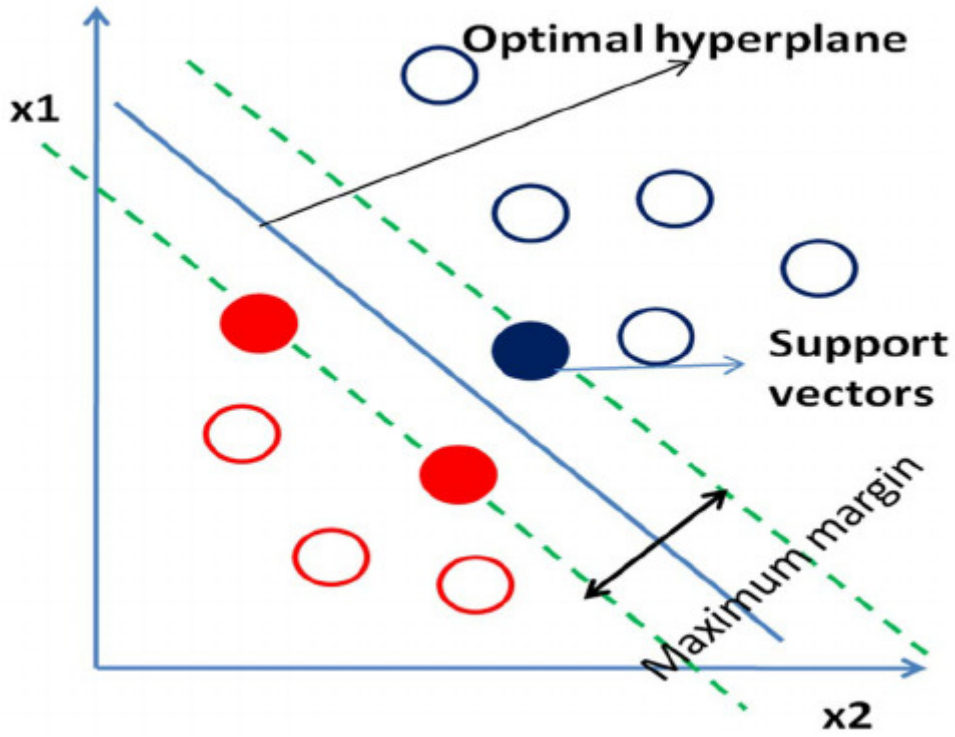
- Clase única vs clase múltiple
  - Clase única (*single-label*)
    - Se asigna una única clase a cada documento
  - Clase múltiple (*multi-label*)
    - Se puede asignar un número cualquiera de clases a cada documento
- Centrado en la clase vs centrado en el documento
  - Centrado en el documento (*document-pivoted*)
    - Dado un documento, encontrar todas las clases a la que pertenece
  - Centrado en la clase (*category-pivoted*)
    - Dada una clase, encontrar los documentos que pertenecen a ella

Clasificación de textos

- Algoritmos
  - Los separadores lineales obtienen buen rendimiento
  - Máquinas de Vectores de Soporte
    - *Support Vector Machines* (SVM)
    - Buen rendimiento con espacios de alta dimensionalidad
    - Lento en construir el modelo pero rápido en clasificar
    - Originalmente para clasificación binaria, pero hay implementaciones para clasificar cuando hay múltiples clases

Clasificación de textos

- Algoritmos
  - Máquinas de Vectores de Soporte



Web Mining


Web
Semantic
search
idea
data
information
results
semantic
practical
article
way
sources
articles
books
published
reading
technical
used
introduction
implementation
blog
make
can
will
really
gives
new
post
explains
examples
Approach
rather
clustering
difficulties
Difficulties
something
literacy
think
works
showing
computer
background
best
Classic
people
meaning
manage
another
words
support
well
subject
part
brain
associations
usability
impression
author
format
users
one
structured
far
matter
many
woven
different
However
behavior
Google
engine
Cluuz
show
use
e.g
i.e
web

Contenidos

- ¿Qué es Web Mining?
- Web Mining vs Text Mining
- Web Content Mining
- Web Structure Mining
- Web Usage Mining

Contenidos

- **¿Qué es Web Mining?**
- Web Mining vs Text Mining
- Web Content Mining
- Web Structure Mining
- Web Usage Mining

¿Qué es Web Mining?

- "Minería Web"
- Aplicación de técnicas de Data Mining para extraer conocimiento a partir de datos en la Web
- Web Mining usa muchas técnicas de Data Mining…
- … pero no es simplemente una aplicación de los métodos tradicionales de Data Mining
  - Heterogeneidad de los datos
  - Naturaleza semi-estructurada o no estructurada

Contenidos

- ¿Qué es Web Mining?
- **Web Mining vs Text Mining**
- Web Content Mining
- Web Structure Mining
- Web Usage Mining

Web Mining vs Text Mining

Text Mining	Web Mining

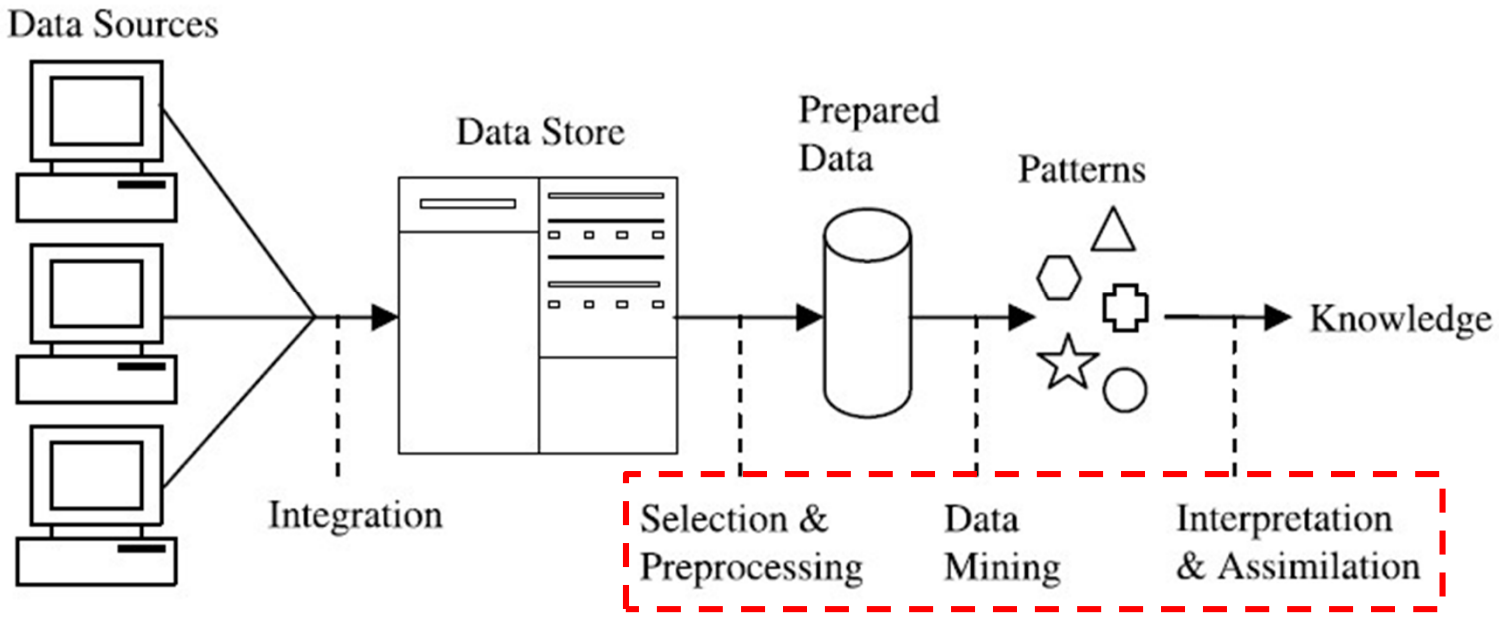
• Datos textuales
  • Formato libre, no estructurados y semi-estructurados
• Dominios
  • Interno/intranet y externo/internet (correos, informes, artículos, …)
• Manejo de los contenidos y organización de la información
	• Datos semi-estructurados
	  • Enlaces y etiquetas HTML
	• Tipos de datos multimedia
	  • Texto, imágenes, audio y vídeo
	• Tanto análisis de contenido como análisis de transacciones

Web Mining vs Text Mining

- En Web Mining, la recolección de datos puede ser una tarea sustancial
- Requiere recuperar un gran número de páginas Web (*crawler*)
- La Web tiene muchas características que la hacen única
  - La cantidad de datos es enorme y sigue creciendo
  - Existen datos de todo tipo
  - La información es heterogénea
  - Una cantidad importante de información está enlazada
  - La información contiene mucho ruido

Web Mining vs Text Mining

- Una vez que se han recolectado los datos, realizamos el mismo proceso en tres pasos



Web Mining vs Text Mining

- Las tareas Web Mining pueden clasificarse en tres tipos
  - Minería del contenido (*Web Content Mining*)
    - Analizar el contenido de la página (texto, imágenes, audio, etc.)
  - Minería de la estructura (*Web Structure Mining*)
    - Analizar la estructura de enlaces (hiperenlaces, etiquetas, etc.)
  - Minería del uso (*Web Usage Mining*)
    - Analizar los datos de uso (logs del servidor, cookies, etc.)

Contenidos

- ¿Qué es Web Mining?
- Web Mining vs Text Mining
- **Web Content Mining**
- Web Structure Mining
- Web Usage Mining

Web Content Mining

- Extrae información útil o conocimiento a partir del contenido de las páginas Web
  - Texto, imágenes, audio y vídeo
- Trabajar con datos textuales implica muchas de las técnicas mencionadas previamente para Text Mining
- Nos vamos a centrar en el procesamiento de contenido textual
  - Trabajar con imágenes, audio y vídeo requiere de herramientas específicas para extraer características

Web Content Mining

- Pre-procesado del contenido
  - Extraer texto del contenido HTML
  - Pasar a minúsculas
  - Obtener la raíz/lema de las palabras
  - Eliminar palabras vacías
  - Selección de características

Web Content Mining

- Creación del vector de características
  - Cada documento (página Web) se representa mediante un vector (disperso) de pesos para cada término
  - Seleccionar un esquema de pesado
    - TF-IDF es el más común
    - Aprovechar la estructura de las páginas Web (cabeceras, tablas, etc.)
      - Ej. dar peso extra a los términos que aparecen en los títulos

Web Content Mining

- Algoritmos
  - Igual que en Data Mining y Text Mining
    - Clasificación
    - Regresión
    - Clustering
    - Reglas de asociación

Web Content Mining

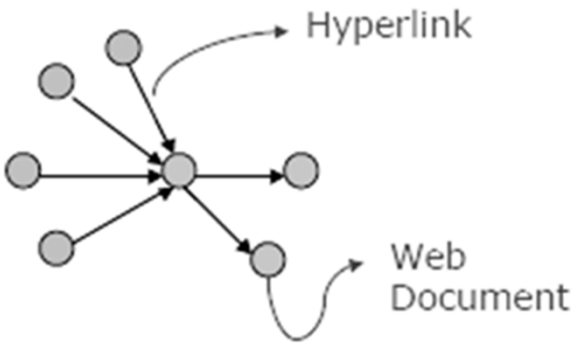
- Aplicaciones
  - Identificar los temas representados en una página Web
  - Clasificar páginas Web
  - Encontrar páginas Web similares en distintos servidores
  - Ordenación de páginas Web (*ranking*)
  - Mostrar/ocultar páginas basados en su relevancia (filtrado)
  - Descubrir patrones en páginas Web para extraer datos útiles (descripción de productos, mensajes en foros, etc.)
  - Analizar críticas de los clientes y mensajes en foros para descubrir las opiniones de los consumidores
  - …

Contenidos

- ¿Qué es Web Mining?
- Web Mining vs Text Mining
- Web Content Mining
- **Web Structure Mining**
- Web Usage Mining

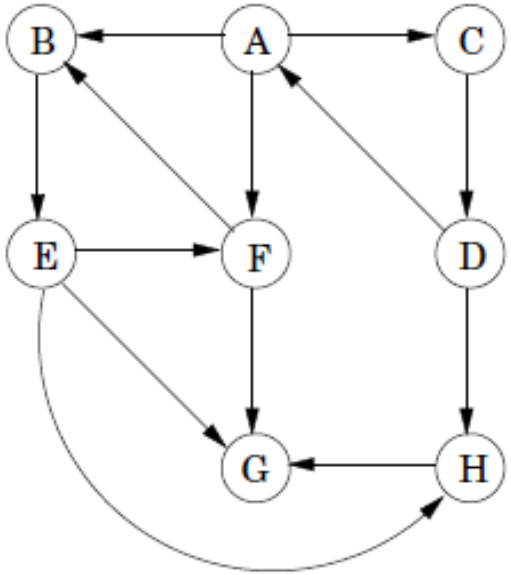
Web Structure Mining

- Descubrir conocimiento útil a partir de los enlaces (*hiperlinks*) existentes en las páginas Web
- Los enlaces tienen dos propósitos principales
  - Navegación
  - Apuntar a páginas con autoridad en el mismo tema que la página que los contiene
- Esto puede usarse para recuperar información útil de la Web



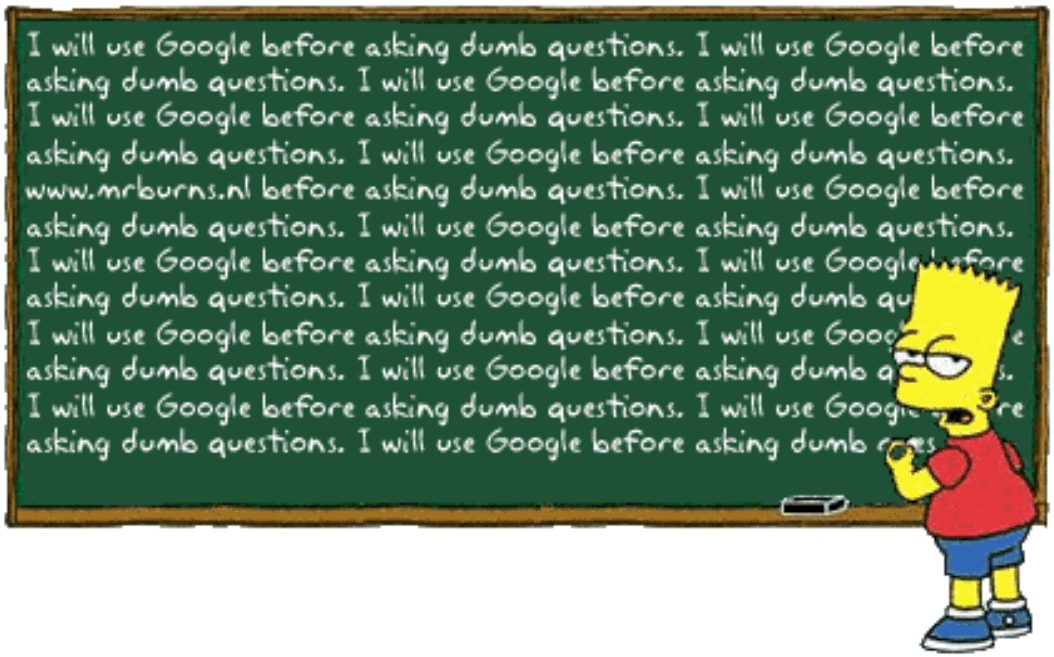
Web Structure Mining

- La Web como grafo
  - Podemos ver la Web, con sus páginas HTML y los enlaces entre ellas, como un grafo dirigido que representa su estructura
    - Cada página web es un nodo
    - Cada enlace es un arco (arista) dirigido



Web Structure Mining

- PageRank
  - El éxito de Google se debe en gran medida a su algoritmo de ordenación basado en enlaces llamado PageRank
  - PageRank tiene su origen en el análisis de redes sociales



Web Structure Mining

- PageRank
  - Es un algoritmo de análisis de enlaces que asigna un peso numérico a cada página Web con el propósito de medir su importancia relativa
  - Hace uso de la estructura de enlaces de la Web para calcular un ranking de calidad para cada página Web
  - No es el único
    - HITS
    - SALSA
    - …

Web Structure Mining

- PageRank
  - Algoritmo simplificado

El PageRank (*PR*) de una página *A* viene dado por:

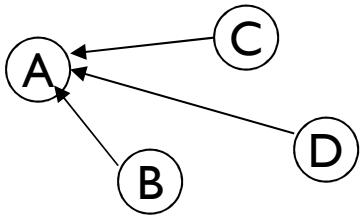
$$PR(A) = \sum_{i=1}^{n} \frac{PR(i)}{L(i)}$$

PR(A) = PageRank de la página *A*
PR(i) = PageRank de todas las páginas *i* que enlazan a *A*
L(i) = número total de enlaces salientes de la página *i* (apunten o no a *A*)

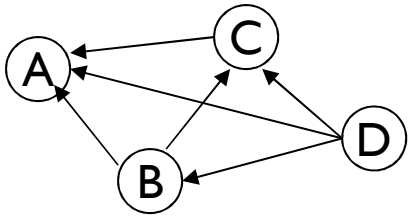
Web Structure Mining

- PageRank
  - Algoritmo simplificado

Asumiendo cuatro páginas: *A*, *B*, *C* and *D*
Imaginemos que cada página comienza con un PageRank de 0.25



$$PR(A) = \frac{PR(B)}{1} + \frac{PR(C)}{1} + \frac{PR(D)}{1} =$$
$$= 0{,}25 + 0{,}25 + 0{,}25 = 0{,}75$$



$$PR(A) = \frac{PR(B)}{2} + \frac{PR(C)}{1} + \frac{PR(D)}{3} =$$
$$= 0{,}125 + 0{,}25 + 0{,}083 = 0{,}458$$

Web Structure Mining

- PageRank
  - Incluyendo factor de amortiguación (*damping factor*)

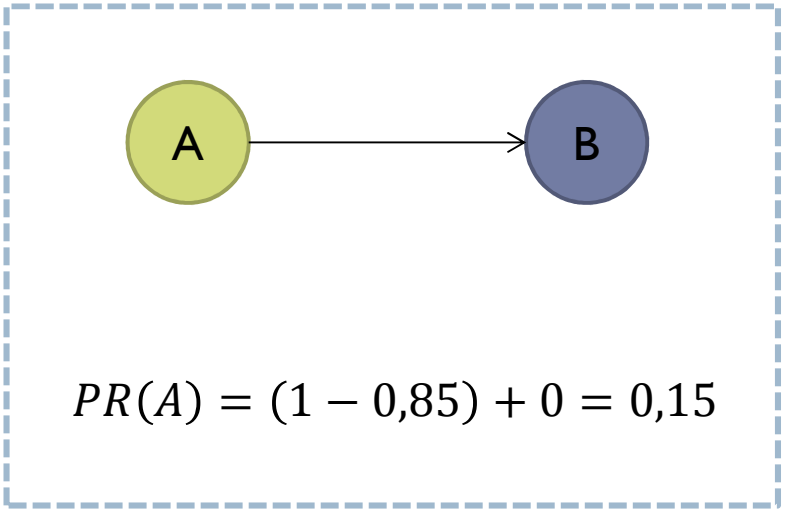
El PageRank de una página *A* viene dado por:

$$PR(A) = (1 - d) + d \sum_{i=1}^{n} \frac{PR(i)}{L(i)}$$

PR(A) = PageRank de la página *A*
PR(i) = PageRank de todas las páginas *i* que enlazan a *A*
L(i) = número total de enlaces salientes de la página *i* (apunten o no a *A*)
d = factor de amortiguación entre 0 y 1 (habitualmente 0,85)

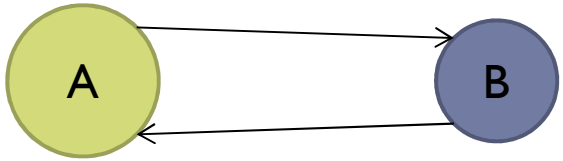
Web Structure Mining

- PageRank
  - Incluyendo factor de amortiguación (*damping factor*)



Web Structure Mining

- PageRank
  - Incluyendo factor de amortiguación (*damping factor*)



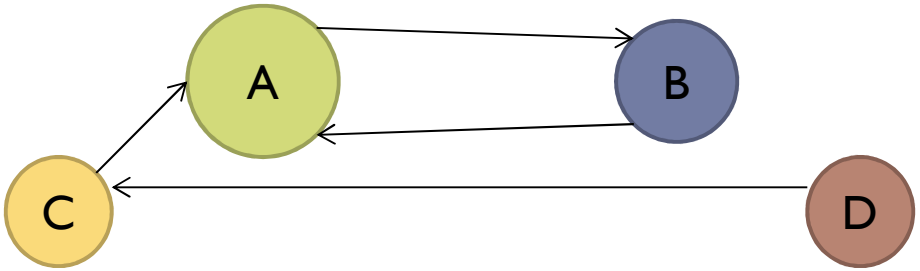
$$PR(A) = (1 - 0{,}85) + 0 = 0{,}15$$

$$PR(B) = (1 - 0{,}85) + 0{,}85\,\frac{0{,}15}{1} = 0{,}278$$

$$PR(A) = (1 - 0{,}85) + 0{,}85\,\frac{0{,}278}{1} = 0{,}1736$$

Web Structure Mining

- PageRank
  - Incluyendo factor de amortiguación (*damping factor*)



$$PR(D) = (1 - 0{,}85) + 0 = 0{,}15$$

$$PR(C) = (1 - 0{,}85) + 0{,}85\,\frac{0{,}15}{1} = 0{,}278$$

$$PR(B) = (1 - 0{,}85) + 0{,}85\,\frac{1{,}736}{1} = 1{,}626$$

$$PR(A) = (1 - 0{,}85) + 0{,}85\left(\frac{1{,}626}{1} + \frac{0{,}15}{1}\right) = 1{,}67$$

Web Structure Mining

- PageRank
  - Justificación intuitiva
    - Un surfista al que se le da una página web de manera aleatoria y va haciendo clic en los enlaces sin volver nunca hacia atrás
    - Al final se aburre y comienza en otra página al azar
    - La probabilidad de que el surfista visite una página es su PageRank
    - Una página puede obtener un PageRank alto
      - Si hay muchas páginas que apuntan a ella
      - Si hay algunas páginas que apuntan a ella y tienen un valor alto de PageRank

Web Structure Mining

- Aplicaciones
  - Calidad de las páginas Web (autoridad y ranking)
  - Identificar estructuras interesantes de la Web (cocitación)
  - Clasificar páginas Web
  - Decidir qué páginas recuperar con un *crawler*
  - Encontrar páginas relacionadas
  - …

Contenidos

- ¿Qué es Web Mining?
- Web Mining vs Text Mining
- Web Content Mining
- Web Structure Mining
- **Web Usage Mining**

Web Usage Mining

- Descubrimiento de patrones de acceso de usuario a partir de los registros (*logs*) de uso de la Web
  - Almacenan cada clic hecho por cada usuario
- Fuentes típicas de datos
  - Datos automáticamente generados en los registros de acceso de los servidores
  - Información sobre las páginas de origen
  - Cookies en el lado del cliente
  - Perfiles de usuario
  - Meta-datos (atributos de la página y el contenido)

Web Usage Mining

- Fichero de registro del servidor Web
  - De aquí es de donde se obtiene la mayoría de información para la tarea de Web Usage Mining

```
218.212.232.82 - - [06/Feb/2005:08:10:11 +0800] "GET
/zencart/images/b_w_grid.gif HTTP/1.1" 200 1128
"http://emefotech.homeunix.net:8000/zencart/" "Mozilla/5.0 (Windows; U; Windows
NT 5.1; en-US; rv:1.7.5) Gecko/20041107 Firefox/1.0"
218.212.232.82 - - [06/Feb/2005:08:10:06 +0800] "GET /zencart/ HTTP/1.1" 200
57249 "-" "Mozilla/5.0 (Windows; U; Windows NT 5.1; en-US; rv:1.7.5)
Gecko/20041107 Firefox/1.0"
218.212.232.82 - - [06/Feb/2005:08:10:14 +0800] "GET
/zencart/images/b_p_grid.gif HTTP/1.1" 200 1165
"http://emefotech.homeunix.net:8000/zencart/" "Mozilla/5.0 (Windows; U; Windows
NT 5.1; en-US; rv:1.7.5) Gecko/20041107 Firefox/1.0“
218.212.232.82 - - [06/Feb/2005:08:10:14 +0800] "GET
/zencart/images/gift_certificates/gv.gif HTTP/1.1" 200 3307
"http://emefotech.homeunix.net:8000/zencart/" "Mozilla/5.0 (Windows; U; Windows
NT 5.1; en-US; rv:1.7.5) Gecko/20041107 Firefox/1.0"
```

Web Usage Mining

- Fichero de registro del servidor Web
  - Una de las claves en Web Usage Mining es el pre-procesamiento de secuencias de clics (*clickstream*)
  - Para producir datos correctos es necesario identificar sesiones
    - Páginas dinámicas
    - Detección y filtrado de robots
    - Identificación de transacciones
      - Identificar usuarios únicos
      - Identificar transacciones únicas de usuarios
    - …

Web Usage Mining

- Aplicaciones
  - Personalización Web
  - Predicción del siguiente evento
  - Descubrimiento de grupos de usuarios con propiedades e intereses comunes
  - Descubrimiento de grupos de usuarios con un comportamiento común
  - Determinar la mejor forma de estructurar un sitio Web
  - Identificar enlaces débiles para su eliminación o mejora
  - …

Bibliography

Copiar de uno es plagio, copiar de dos es investigación

Wilson Mizner

Bibliography

- *Automatic Text Classification*, Yutaka Sasaki (2008)
- *Introduction to Text Mining*, Mandar Mitra (2014)
- *Machine Learning in Automated Text Categorization*, Fabrizio Sebastiani (2002)
- *Mining Text Data*, Char C. Aggarwal and ChengXiang Zhai (2012)
- *Natural Language Processing*, Johanna Moore (2011)
- *Natural Language Processing*, Joyce Chai (2011)
- *Principles of Data Mining*, Max Bramer (2013)
- *Statistical Methods in Lexical Semantics*, Karin Verspoor (2010)

Bibliography

- *The Anatomy of a Large-Scale Hypertextual Web Search Engine*, Sergey Brin and Lawrence Page (1998)
- *The Unreasonable Effectiveness of Data*, Alon Halevy et al. (2009)
- *Topics in Statistical Semantics*, Desislava Zhekova & Hinrich Schütze (2013)
- *Text and Web Mining*, Nguyen Hung Son (2000)
- *Web Data Mining*, Bing Liu (2007)
- *Web Mining: Accomplishments & Future Directions*, Jaideep Srivastava (2003)


That's All
Folks