Análisis y Recuperación de Información

1er Cuatrimestre 2017

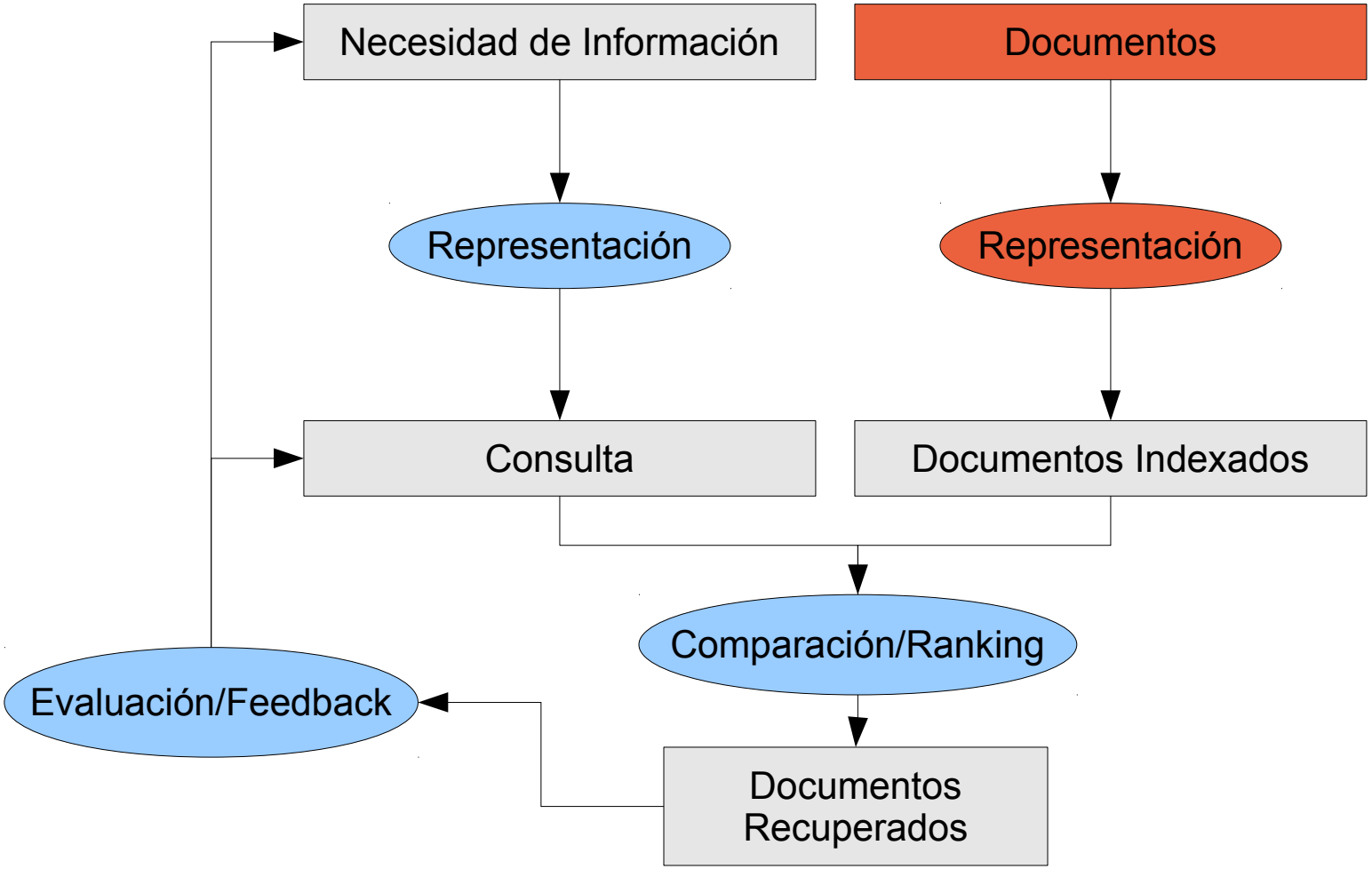
Página Web

http://www.exa.unicen.edu.ar/catedras/ayrdatos/

Prof. Dra. Daniela Godoy

ISISTAN Research Institute
UNICEN University
Tandil, Bs. As., Argentina
http://www.exa.unicen.edu.ar/~dgodoy
dgodoy@exa.unicen.edu.ar

Recuperación de Información



Pre-procesamiento

- La representación del contenido de documentos es la transformación automática del texto en una forma que represente uno o más aspectos de su significado
- Se basa en técnicas:
  - estadísticas
  - lingüísticas
  - basadas en conocimiento

Pre-procesamiento

- El pre-procesamiento del texto consiste en general en los siguientes pasos:
  - reducción a un formato ASCII (eliminación de caracteres de formato, estilo, etc.)
  - conversión a minúscula
  - identificación de palabras del texto (strings de caracteres contiguos delimitados por blancos)
  - eliminación de puntuación
  - eliminación de Stop-Words
  - stemming
  - asignación de pesos a términos

Tokenización

- Entrada: “retrieval, organization and storage”
- Salida: tokens
  - retrieval
  - organization
  - and
  - storage
- Cada token es candidato para ser una entrada en el índice, luego de algún procesamiento

Tokenización

- Problemas:
  - acentuación: *résumé/resume*
  - apóstrofes: *L'ensemble*
  - abreviaciones: *U.S.A./USA*
  - cómo acostumbran los usuarios a escribir las consultas para estas palabras?
- Reducción a letras minúsculas
  - excepción: letras mayúsculas en el medio de una frase
    - ejemplo: *General Motors*

Eliminación de Stop-words

- Stop-words son palabras que por su frecuencia y/o semántica no poseen valor discriminatorio alguno, es decir no permiten distinguir un documento de otro en una colección
- Habitualmente se trata de artículos, pronombres, preposiciones, verbos muy frecuentes, adverbios, etc.

Eliminación de Stop-words

- Efectos negativos:
  - su alta frecuencia hace que cualquier función de asignación de pesos tienda a disminuir el impacto del resto de las palabras en el documento
  - insumen gran cantidad de tiempo de procesamiento improductivo
- Efectos positivos:
  - su eliminación reduce en más de un 30% el tamaño del documento

Eliminación de Stop-words

- La eliminación de stop-words se realiza chequeando el contenido del documento contra un listado disponible
- Las listas de stop-words pueden ser:
  - independientes de la colección, cada lenguaje posee listas estándares de stop-words de longitud variable
  - dependientes de la colección, palabras que para una determinada colección no poseen valor discriminante (por ejemplo, en computación la palabras “software”)

Eliminación de Stop-words

a	also	appreciate	becoming	besides
able	although	appropriate	been	best
about	always	are	before	better
above	am	around	awfully	between
according	among	as	b	beyond
accordingly	amongst	aside	be	both
across	an	ask	became	brief
actually	and	asking	because	but
after	another	associated	become	by
afterwards	any	at	becomes	...
again	anybody	available	becoming	
against	anyhow	Away	been	
all	anyone	awfully	before	
allow	anything	b	beforehand	
allows	anyway	be	behind	
almost	anyways	became	being	
alone	anywhere	because	believe	
along	apart	become	below	
already	appear	becomes	beside	

Eliminación de Stop-words

- Texto Original:
  - Information Systems Asia Web - provides research, IS-related commercial materials, interaction, and even research sponsorship by interested corporations with a focus on Asia Pacific region
  - Survey of Information Retrieval - guide to IR, with an emphasis on web-based projects. Includes a glossary, and pointers to interesting papers
- Texto resultante al eliminar stop-words:
  - Information Systems Asia Web provides research IS-related commercial materials interaction research sponsorship interested corporations focus Asia Pacific region
  - Survey Information Retrieval guide IR emphasis web-based projects Includes glossary pointers interesting papers

Stemming

- Un algoritmo de stemming es un proceso de normalización lingüística en el cual las diferentes formas que puede adoptar una palabra son reducidos a una única forma común, a la cual se denomina stem
  - computer, computers, compute, computes, computational, computationally, etc.
  - comput
- El stem conlleva el significado del concepto asociado a un grupo de palabras
- Efectos positivos:
  - mejora la formulación de consultas (incrementa el recall)
  - reduce la dimensión del espacio de términos (10% y 50%)

Stemming

- Se cuenta con un diccionario que posee el *stem* asociado a cada palabra

TERMINO	STEM
engineering	engineer
engineered	engineer
engineer	engineer

- Usualmente se emplea este método en conjunción con la eliminación de sufijos

Stemming

- Algoritmos que eliminan sufijos y/o prefijos
  - Algoritmo de Harman
    - Plural a singular
    - Tercera persona a primera persona
  - Algoritmo de Lovins
    - 260 sufijos
    - Sufijos de mayor coincidencia
  - Algoritmo de Porter
    - 60 sufijos en diferentes grupos
    - Se aplican los sufijos de un grupo antes de pasar al siguiente

Stemming

PASO	CONDICION	SUFIJO		EJEMPLO
1a	NULL	sses	ss	stresses -> stress
	NULL	ies	I	ponies -> poni
	NULL	ss	ss	caress -> caress
	NULL	s	NULL	cats -> cat
1b	*v*	ing	NULL	making -> make
	. . .	. . .	. . .	. . .
1b1	NULL	at	ate	inflat(ed) -> inflaste
	. . .	. . .	. . .	. . .
1c	*v*	y	I	happy -> happi
2	m > 0	aliti	al	formaliti > formal
	m > 0	izer	ize	digitizer -> digitize
	. . .	. . .	. . .	. . .
3	m > 0	icate	ic	duplicate -> duplic
	. . .	. . .	. . .	. . .
4	m > 1	able	NULL	adjustable -> adjust
	m > 1	icate	NULL	microscopic -> microscop
	. . .	. . .	. . .	. . .
5a	m > 1	e	NULL	inflate -> inflat
	. . .	. . .	. . .	. . .
5b	m > 1, *d, *<L>	NULL	single letter	controll -> control, roll -> roll

m es número de veces que se repite una secuencia vocal-consonante
v el stem contiene al menos una vocal
*d el stem termina en doble consonante

Stemming

- Texto Original:
  - marketing strategies carried out by U.S. companies for their agricultural chemicals, report predictions for market share of such chemicals, or report market statistics for agrochemicals, pesticide, herbicide, fungicide, insecticide, fertilizer, predicted sales, market share, stimulate demand, price cut, volume of sales
- Texto resultante de aplicar el algoritmo de Porter:
  - market strateg carr compan agricultur chemic report predict market share chemic report market statist agrochem pesticid herbicid fungicid insecticid fertil predict sale stimul demand price cut volum sale

Stemming

- Un algoritmo de stemming puede producir resultados incorrectos ya sea por under-stemming o over-stemming
  - *Over-stemming*: términos con diferente significado son transformados a un mismo stem. Por ejemplo:
    - “policy”/“police”, “university”/“universe”, “organization”/“organ”
  - *Under-stemming*: términos con similar significado no son reducidos a una misma raíz. Por ejemplo:
    - “European”/“Europe”, “matrices”/“matrix”,“machine”/“machinery”
- Over-stemming reduce la precisión del sistema mientras que under-stemming su recall

Relaciones Semánticas

- Se basa en el empleo de diccionario para detectar relaciones entre palabras, tales como:
  - Sinonimia (Synonymy)
  - Polisemia (Polysemy) / Antonimia (Antonymy)
  - Hiponimia/Hipernimia (Hyponymy/Hypernym)
  - Meronimia (Meronymy) / Holonimia (Holonymy)
- Los términos en el documento son trasladados a conceptos de mayor nivel de abstracción

Relaciones Semánticas

- Sinonimia
  - Diferentes formas de expresar conceptos relacionados
- Polisemia
  - Homonimia: la misma palabra con diferente significado
    - bank (river)
    - bank (money)
  - Polisemia: diferentes sentidos de la misma palabra
- Antonimia
  - Palabras con semántica opuesta:
    - antonym(large, small)
    - antonym(big, small)
    - antonym(big, little)

Relaciones Semánticas

- Hiponimia/Hipernimia
  - Relaciones del tipo es un
    - hyponym(robin,bird)
    - hyponym(bird,animal)
    - hyponym(emu,bird)
  - A es un es a hipérnimo de B si B es un tipo de A
  - A is a hipónimo de B si A es un tipo de B
- Meronimia/Holonimia
  - Relaciones del tipo parte de
    - part of(beak, bird)
    - part of(bark, tree)

Relaciones Semánticas

- Un diccionario es un conjunto de palabras sobre el cual se define una topología que subraya las interrelaciones semánticas entre ellas
- Un diccionario puede ser:
  - Temático: definido para una disciplina en un idioma determinado
  - De propósito general: definido para todo un idioma

Relaciones Semánticas

- Base de datos léxica para el lenguaje inglés
  - WordNet
    - http://www.cogsci.princeton.edu/~wn/main/

    WordNet on the WWW
    - http://www.cs.buffalo.edu/~aec/wordnet-start.html

- EuroWordNet es un proyecto de desarrollar una base de datos conteniendo relaciones semánticas entre palabras para varios lenguajes europeos (Holandés, Italiano y Español)
  - The EuroWordNet Project
    - http://www.hum.uva.nl/~ewn/
    - http://www.dcs.shef.ac.uk/research/groups/nlp/funded/eurowordnet.html

Zipf's Law

- Unas pocas palabras son muy comunes
  - 2 de las palabras más frecuentes (por ejemplo, “the” y “and”) son alrededor del 10% de las ocurrencias de las palabras
- Muchas palabras son raras
  - la mitad de las palabras de una colección ocurren una única vez

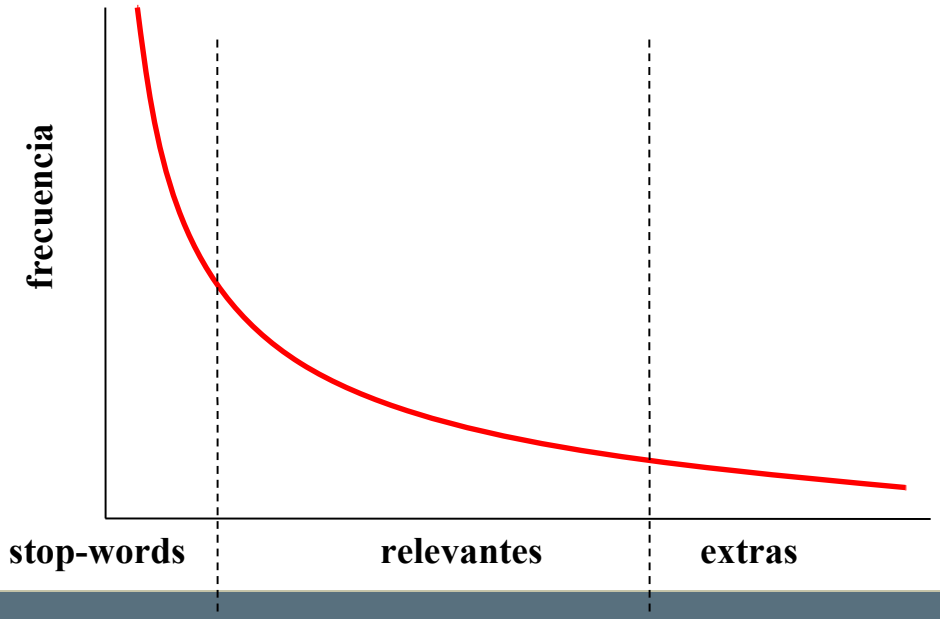
Zipf's Law

Frequent Word	Number of Occurrences	Percentage of Total
the	7,398,934	5.9
of	3,893,790	3.1
to	3,364,653	2.7
and	3,320,687	2.6
in	2,311,785	1.8
is	1,559,147	1.2
for	1,313,561	1.0
The	1,144,860	0.9
that	1,066,503	0.8
said	1,027,713	0.8

Frequencies from 336,310 documents in the 1GB TREC Volume 3 Corpus
125,720,891 total word occurrences; 508,209 unique words

Zipf's Law

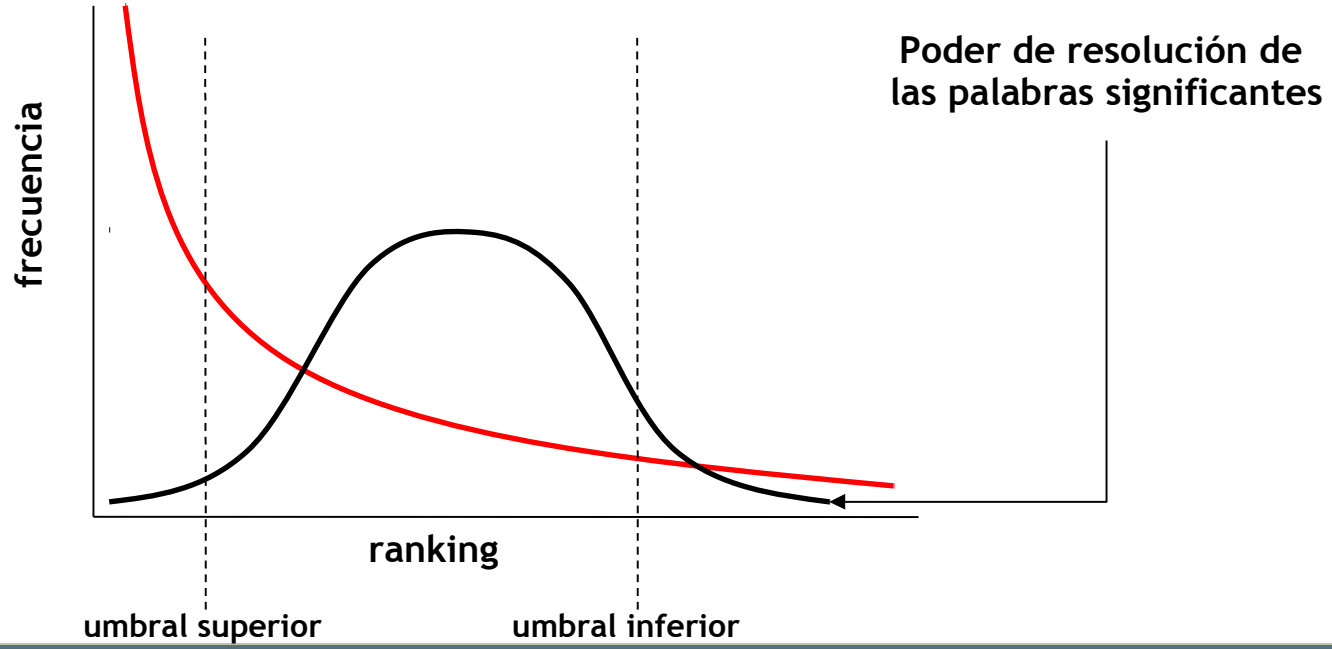
- Las palabras en un texto no se encuentran uniformemente distribuidas
- La ocurrencia de palabras no posee una distribución normal sino que exhibe una distribución denominada Zipf
- Poder de resolución: la habilidad de las palabras de discriminar contenidos



Zipf's Law

El producto de la frecuencia de palabras (f) y su ranking (r) es aproximadamente constante, siendo el ranking el orden de las palabras por frecuencia de ocurrencia

- unos pocos elementos ocurren con muy alta frecuencia
- el número medio de elementos ocurre con una frecuencia media
- muchos elementos ocurren muy infrecuentemente



Zipf's Law

- Efectos positivos:
  - stop-words son una gran fracción del texto de manera que eliminarlas reduce el costo de almacenamiento del archivo invertido
  - la lista de postings (ocurrencias) de las palabras restantes va a ser corta ya que son raras, agilizando la recuperación
- Efecto negativo:
  - para algunas palabras extraer suficiente datos para hacer un análisis estadístico significante (por ejemplo, correlación para expansión de consultas) es dificultoso porque son raras

N-grams

- Técnica de generación de frases basada en frecuencia de n-grams
  - N-gram es una secuencia de n palabras consecutivas (ejemplo, “microsoft windows” es un 2-gram, “Word for Windows” es un 3-gram),
  - N-grams frecuentes son los n-grams que aparecen en la colección con una frecuencia mínima MinFreq
- N-grams es una técnica interesante por la simplicidad y eficiencia del algoritmo

N-grams

- Dado:
  - un conjunto de documentos (cada documento es una secuencia de palabras)
  - MinFreq, la frecuencia mínima para n-grams
  - MaxNGramSize, la longitud máxima de n-grams

- For Len=1 to MaxNGramSize do
  - generar n-grams candidatos con secuencias de palabras de tamaño Len usando las n-grams frecuentes de longitud Len-1
  - borrar n-grams candidatos con frecuencia menos que MinFreq

Extracción de Información

- Reconocimiento de Nombres de Entidades
  - identificación de elementos en el texto dentro de un conjunto de categorías predefinidas
    - tres categorías aceptadas comunmente: personas, lugares y organizaciones
    - Otros elementos comunes: fechas, medidas (porcentages, valores monetarios, pero, etc.), direcciones de mail, etc.
    - en dominios específicos: nombre de drogas, condiciones médicas, referencias bibliográficas, etc.

- Metadatos como #COMPANY o #PERSON se agregan al indexar

Extracción de Información

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying…

Extracción de Información

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying…

Microsoft Corporation
CEO
Bill Gates
Microsoft
Gates
Microsoft
Bill Veghte
Microsoft
VP
Richard Stallman
founder
Free Software Foundation

Extracción de Información

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying…

Microsoft Corporation
CEO
Bill Gates

Microsoft
Gates

Microsoft

Bill Veghte
Microsoft
VP

Richard Stallman
founder
Free Software Foundation

Extracción de Información

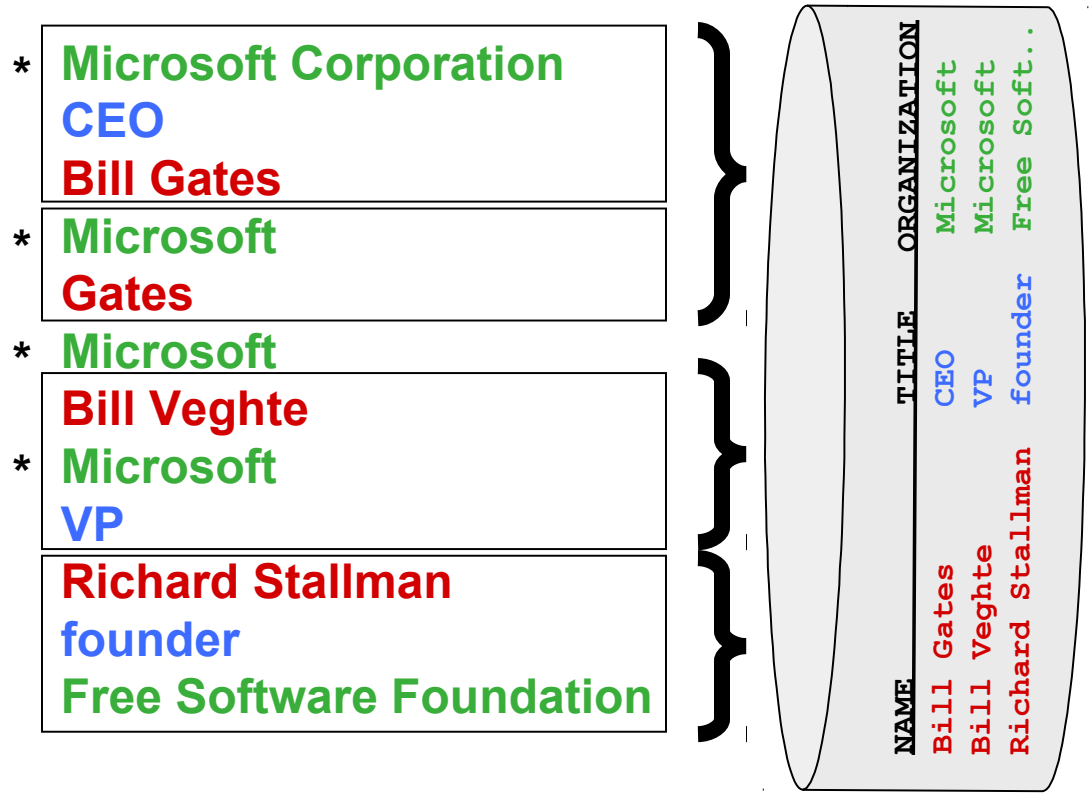
October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access."

Richard Stallman, founder of the Free Software Foundation, countered saying...



Extracción de Información

- Técnicas basadas en gramática
  - hand-crafted
  - requiere mucho trabajo manual y mantenimiento
  - no es portable
  - mayor exactitud
- Modelos estadísticos
  - requieren gran cantidad de datos anotados
  - son portables

Extracción de Información

- Shallow Parsing
  - Evidencia interna: los nombres usualmente tienen una estructura interna

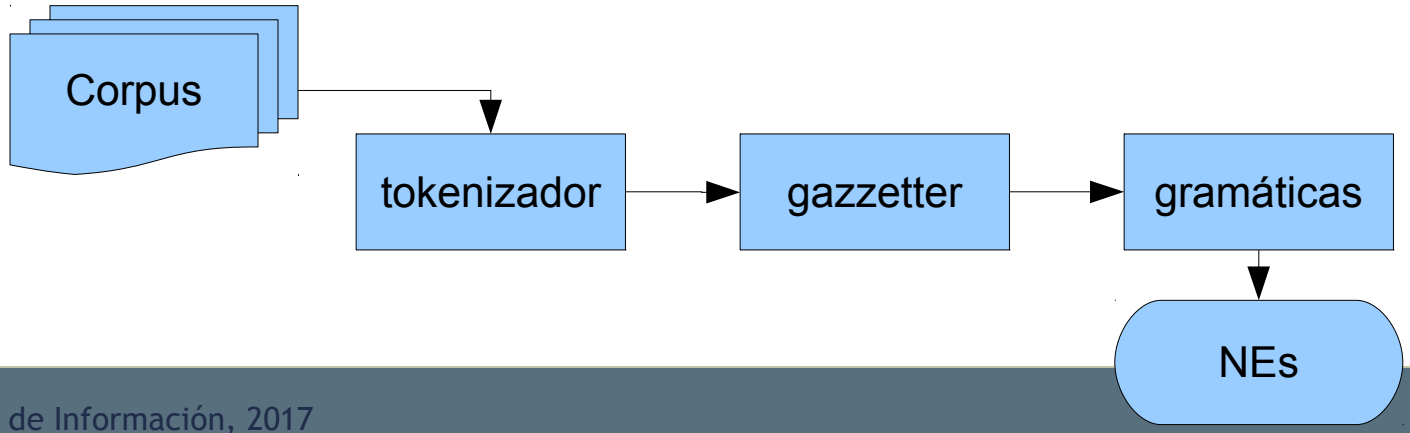
    location:
    Cap Word + {Street, Boulevard, Avenue, Crescent, Road}
    e.g. Portobello Street

  - Evidencia externa: los nombres se usan en un contexto predecible

    location:
    "to the" COMPASS "of" CapWord
    e.g. to the south of Loitokitok

Extracción de Información

- Tokenizador
- Gazetteer lists
  - NEs: ciudades, nombres, países, etc.
  - palabras claves: por ej. designadores de companías (S.A., Co.), títulos (Sr., Ms., PhD, Dr.), etc.
  - prefijos, sufijos, etc.
- Gramáticas
  - reglas definidas manualmente para reconocimiento de NEs



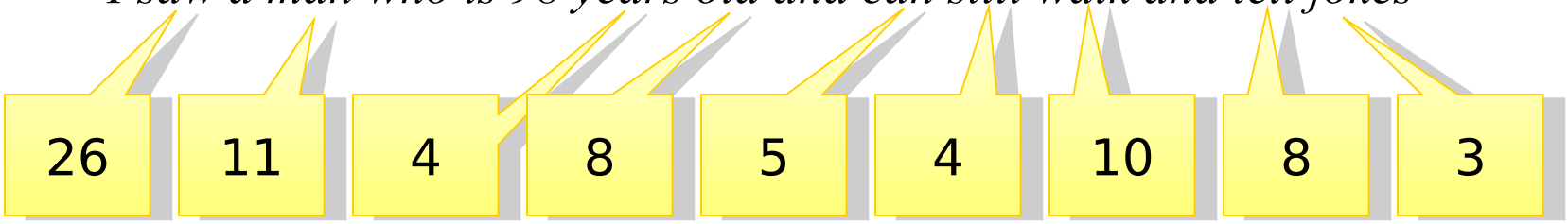
Extracción de Información

- Texto Original:
  - John Davenport, 52 years old, was appointed chief executive office of this international telecommunications concern's U.S. subsidiary, Cable & Wireless North America Inc.. Mr. Davenport, who succeeds Jon Zrno, is currently general manager of the group's operations in Bermuda.
- Una forma de indexar:
  - john davenport year old appoint chief executive office international telecommunication concern subsidiary cable wireless north america davenport succeed jon zrno current general manager group operation bermuda
- Otra forma de indexar con NEs:
  - John_Davenport #PERSON 52 years_old #AGE appoint chief_executive_office #JOB international telecommunication concern U.S. #COUNTRY subsidiary Cable_&_Wireless_North_America #COMPANY Davenport #PERSON succeed Jon_Zrno #PERSON current general_manager #JOB group operation Bermuda #COUNTRY

Desambiguación lingüística

- Word Sense Disambiguation (WSD)
  - desambiguación es el problema de seleccionar el sentido de una la palabra dentro de un conjunto predefinido de posibilidades

I saw a man who is 98 years old and can still walk and tell jokes



43,929,600
sentidos

Desambiguación lingüística

- Enfoque basado en conocimiento
  - uso de recursos léxicos externos como diccionarios y tesauros
- Enfoque supervisado
  - usado con un conjunto de entrenamiento
  - clasificación de una entrada (con contexto) y las categorías son los sentidos correctos
- Enfoque no supervisado
  - basado en un conjunto de ejemplos sin un sentido asignado

Desambiguación lingüística

- Enfoque basado en conocimiento
  - Para cada palabra en el vocabulario el diccionario provee:
    - Una lista de significados
    - Definiciones
    - Ejemplos típicos de uso

WordNet definiciones para el sustantivo “plant”

- buildings for carrying on industrial labor; "they built a large plant to manufacture automobiles“
- a living organism lacking the power of locomotion
- something planted secretly for discovery by another; "the police used a plant to trick the thieves"; "he claimed that the evidence against him was a plant"
- an actor situated in the audience whose acting is rehearsed but seems spontaneous to the audience

Desambiguación lingüística

- Un tesauro agrega
  - Relaciones de sinonímia entre significados

WordNet synsets for the noun “plant”

1. plant, works, industrial plant
2. plant, flora, plant life

Desambiguación lingüística

- Una red semántica puede agregar:
  - Hipernimia/hiponimia (IS-A), meronimia/holonimia (PART-OF), etc.

Word Net conceptos relacionados a “plant life”
{plant, flora, plant life}
hypernym: {organism, being}
hypomym: {house plant}, {fungus}, …
meronym: {plant tissue}, {plant part}
holonym: {Plantae, kingdom Plantae, plant kingdom}

Desambiguación lingüística

- Algoritmo de Lesk (1986)
  - identifica sentidos en un contexto dado usando superposición
  - simultáneamente para todas las palabras en un contexto
- Algoritmo:
  - recupera del diccionario todas las definiciones para los sentidos de las palabras a ser desambiguadas
  - determina la superposición de definiciones para todas las combinaciones posibles de sentidos
  - elige los sentidos con mayor grado de superposición

Desambiguación lingüística

- Desambiguar PINE CONE

PINE
1. kinds of evergreen tree with needle-shaped leaves
2. waste away through sorrow or illness

CONE
1. solid body which narrows to a point
2. something of this shape whether solid or hollow
3. fruit of certain evergreen trees

Pine#1 ∩ Cone#1 = 0
Pine#2 ∩ Cone#1 = 0
Pine#1 ∩ Cone#2 = 1
Pine#2 ∩ Cone#2 = 0
Pine#1 ∩ Cone#3 = 2
Pine#2 ∩ Cone#3 = 0

Desambiguación lingüística

- Enfoque supervisado
  - métodos que inducen un clasificador a partir de un conjunto de texto tageado manualmente
- Requiere:
  - texto anotado con los sentidos correctos
  - análisis sintáctico (POS tagger)
- Objetivo
  - destinado a desambiguar una única palabra
  - se selecciona un conjunto de características para representar el contexto, se transforma en un vector
  - clasifica una palabra en el sentido más apropiado en base a su contexto (co-ocurrencias, POS tags, relaciones verbo-objetos, etc.)

Desambiguación lingüística

- *An electric guitar and bass player stand off to one side, not really part of the scene, just as a sort of nod to gringo expectations perhaps*

- Contexto vecino (características locales)
  - [(guitar, NN1), (and, CJC), (player, NN1), (stand, VVB)]
- Palabras que co-ocurren frecuentemente:
  - [fishing, big, sound, player, fly, rod, pound, double, runs, playing, guitar, band]
  - [0,0,0,1,0,0,0,0,0,0,1,0]
- Otras características:
  - [followed by "player", contains "show" in the sentence,…]
  - [yes, no, …]

Desambiguación lingüística

- Enfoque no supervisado
  - métodos que agrupan palabras en base a la similitud de su contexto
  - hipótesis de contexto fuerte:
    - las palabras con significados similares tienden a ocurrir en contextos similares (Miller and Charles, 1991)
- Requiere:
  - gran corpus de textos
  - sólo usa el texto, sin recursos externos o anotaciones
- Objetivo
  - destinado a desambiguar una única palabra en un contexto
  - mide similitud entre contextos

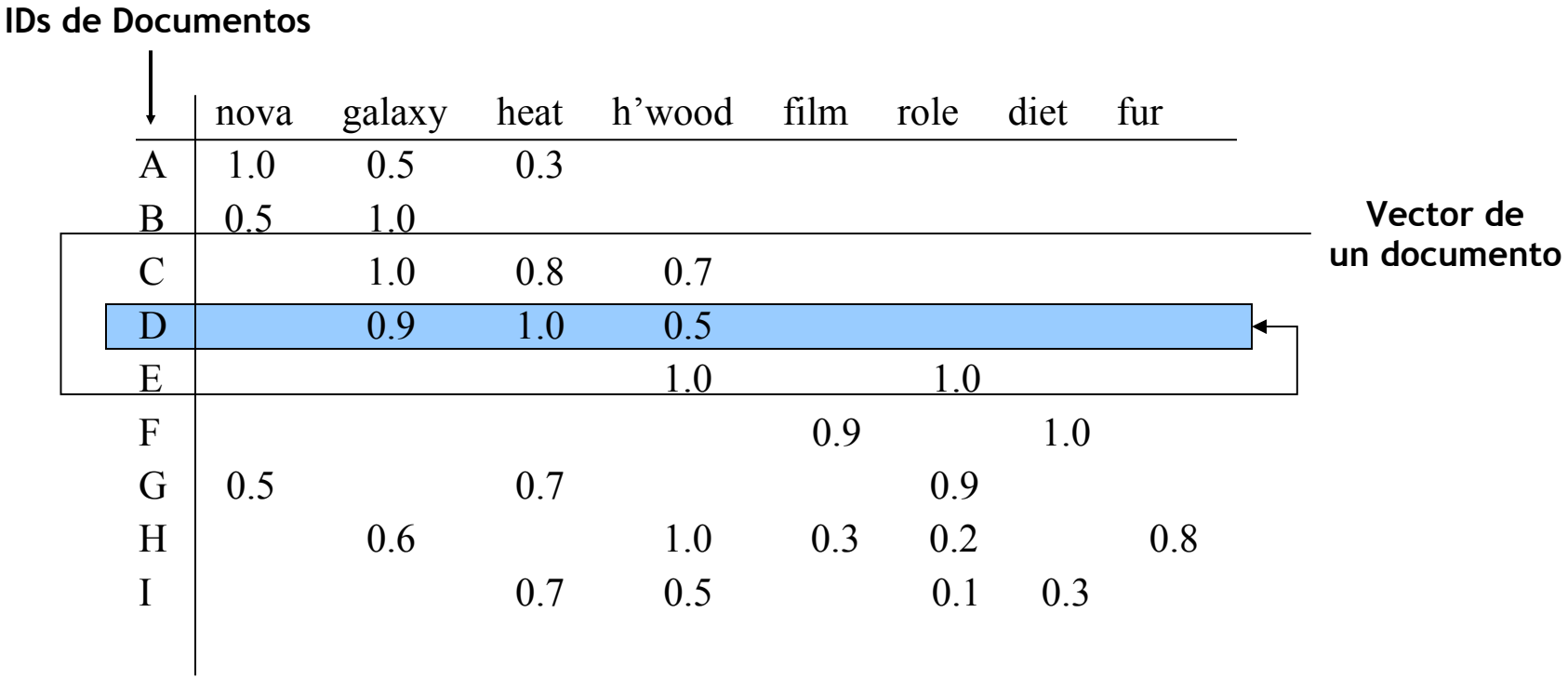
Desambiguación lingüística

- las características pueden identificarse en datos de entrenamiento o en los datos a ser agrupados
- no asigna sentidos o etiquetas a los grupos
- encuentra las palabras que ocurren en contextos similares y las agrupan

Representación de Documentos

- Los documentos se representan usualmente como “bags of words”, cuya representación computacional es a través de vectores
- En este modelo los documentos se mapean a un espacio de vectores altamente dimensional
  - cada documento consiste en una secuencia de términos
  - los términos únicos en un conjunto de documentos determinan las dimensiones del espacio
- Representaciones más sofisticadas:
  - parecen tener mejor calidad
  - los experimentos conducidos no conducen a una mejora significante sobre los enfoques basados en términos

Pesado de Términos



Pesado de Términos

- Los vectores incluyen sólo la presencia (1) o la ausencia (0) de un término

Docs	*t1*	*t2*	*t3*
D1	1	0	1
D2	1	0	0
D3	0	1	1
D4	1	0	0
D5	1	1	1
D6	1	1	0
D7	0	1	0
D8	0	1	0
D9	0	0	1
D10	0	1	1
D11	1	0	1
Q	1	1	1

Pesado de Términos

- Los términos más frecuentes es un documento son los más importantes o más indicativos del tema del documento

f_{ik} frecuencia del término i en el documento k

- Se puede normalizar la frecuencia de un término fik dividiéndola por la frecuencia del término más común en el documento

$$tf_{ik} = \frac{f_{ik}}{\max f_{ik}}$$

Pesado de Términos

- El vector incluye la frecuencia de ocurrencia de cada término

Docs	*t1*	*t2*	*t3*
D1	2	0	3
D2	1	0	0
D3	0	4	7
D4	3	0	0
D5	1	6	3
D6	3	5	0
D7	0	8	0
D8	0	10	0
D9	0	0	1
D10	0	3	5
D11	4	0	1
Q	1	2	3

Pesado de Términos

- TF-IDF mide:
  - frecuencia del término (TF – term frequency)
  - frecuencia inversa de documentos (IDF – inverse document frequency)
- Se desea dar mayor peso a los términos que:
  - son frecuentes en los documentos relevantes... PERO
  - son infrecuentes en la colección como un todo
- Se asigna un peso TF x IDF a cada término en cada documento

Pesado de Términos

$$w_{ik} = tf_{ik} * idf_k$$

t_k = término k del documento D_i

tf_{ik} = frecuencia del término t_k en el documento D_i

idf_k = frecuencia inversa de documentos del término t_k en C

N = número total de documentos en la colección C

n_k = número de documentos en C que contienen a t_k

$$idf_k = \log\left(\frac{N}{n_k}\right)$$

Pesado de Términos

- La frecuencia inversa de documentos (IDF) provee valores altos para palabras raras y bajos para palabras comunes

$$\log\left(\frac{10000}{10000}\right)=0$$

$$\log\left(\frac{10000}{5000}\right)=0.301$$

$$\log\left(\frac{10000}{20}\right)=2.698$$

$$\log\left(\frac{10000}{1}\right)=4$$