Representación de Documentos

AGENTES INTELIGENTES EN INTERNET

Daniela Godoy - ISISTAN

Representación de Documentos

- El preprocesamiento 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 las Palabras

Representación de Documentos

- Análisis de contenido es la tranformación automática del texto en una forma que represente uno o más aspectos de su significado.
- Se basa en técnicas:
  - Estadísticas
  - Lingüisticas
  - Basadas en Conocimiento

Representación de Documentos
Eliminación de Stop-Words

- Stop-words son palabras que por su frecuenciay/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.

Representación de Documentos
Eliminación de Stop-Words

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

Representación de Documentos
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 estándars de stop-words de longitud variada
  - Dependientes de la colección, palabras que para una determinada colección no poseen valor discriminante (por ej. en computación la palabras “software”)

Representación de Documentos
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	

Representación de Documentos
Stemming

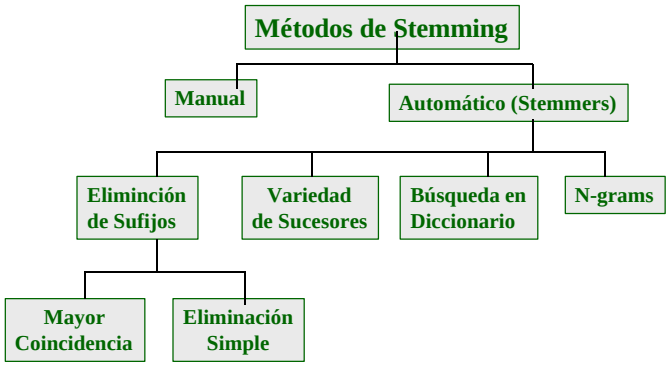
- Un algoritmo de stemming es un proceso de normalización lingüistica en el cual las diferentes formas que puede adoptar una palabra son reducidos a una única forma común, a la cual se denomina stem

  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 terminos (10% y 50%)

Representación de Documentos
Stemming



Representación de Documentos
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

Representación de Documentos
Stemming

- Algoritmos que eliminan sufijos y/o prefijos
- Algunos algoritmos disponibles son:
  - 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

Representación de Documentos
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

Representación de Documentos
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

Representación de Documentos
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 una misma raíz. 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*

Representación de Documentos
Relaciones Semánticas

- Se basa en el empleo de diccionario para detectar relaciones entre palabras, tales como:
  - Sinonimia (Synonymy)
  - Metonimia (Metonymy)
  - Hipofísis/Hiperfísis (Hyponymy/Hyperonymy)
  - Meronimia (Meronymy)
  - Antonimia (Antonymy)
- Los términos en el documento son translados a conceptos de mayor nivel de abstracción

Representación de Documentos
Synonymy, Polysemy y Antonymy

- Synonymy
  - Diferentes formas de expresar conceptos relacionados
- Polysemy
  - Homonym: la misma palabra con diferente significado
    - bank (river)
    - bank (money)
  - Polysem: diferentes sentidos de la misma palabra
- Antonymy
  - Palabras con semántica opuesta:
    - antonym(large, small)
    - antonym(big, small)
    - antonym(big, little)

Representación de Documentos
Hyponymy, Hyperonymy y Meronymy

- Hyponymy/Hyperonymy
  - 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
- Meronymy
  - Relaciones del tipo parte de
    - part of(beak, bird)
    - part of(bark, tree)

Representación de Documentos
Diccionarios

- 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

Representación de Documentos
WordNet

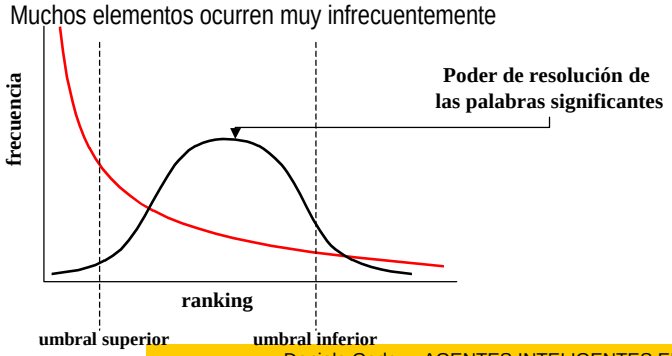
- Base de datos léxica para el lenguaje inglés
  - WordNet
    - http://www.cogsci.princeton.edu/~wn/main/
  - WordNet on the WWW
    - http://www.cs.buffalo.edu/~aec/wordnet-start.html
- EutoWordNet es un proyecto de desarrollar una base de datos conteniendo relaciones semánticas entre palabras para 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

Representación de Documentos
Propiedades Estadísticas del Texto

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

Representación de Documentos
Zipf 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



Representación de Documentos
Modelo de Espacio de Vectores

- Los documentos se representan usualmente como “bags of words”, cuya representación computacional es a través de vectores

Ids de Dcoumentos

	nova	galaxy	heat	h'wood	film	role	diet	fur
A	1.0	0.5	0.3					
B	0.5	1.0						
C		1.0	0.8	0.7				
D		0.9	1.0	0.5				
E				1.0		1.0		
F					0.9		1.0	
G	0.5		0.7			0.9		
H		0.6		1.0	0.3	0.2		0.8
I			0.7	0.5		0.1	0.3	

Un vector de documento

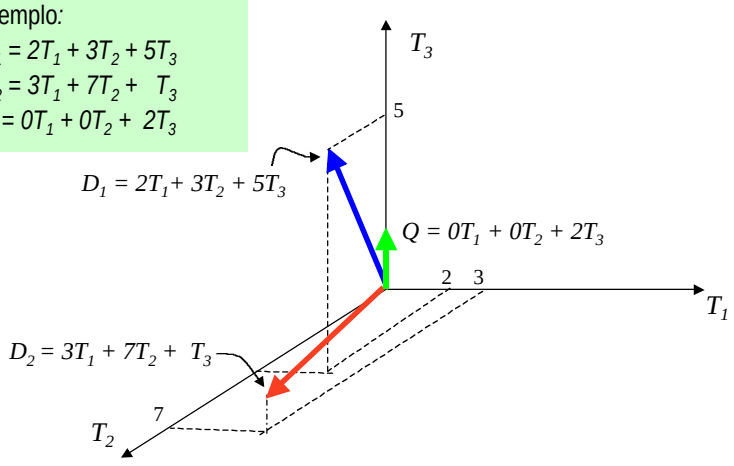
Representación de Documentos
Modelo de Espacio de Vectores

- Modelo de Espacio de Vectores
  - Asume t términos distintos después del preprocesamiento, el vocabulario
  - Estos términos conforman un espacio de vectores Dimension = t = |vocabulario|
  - Cada término, i, en un documento o consulta, j, tiene un peso, w_{ij}.
  - Ambos, documentos y consultas, se expresan como vectores t-dimensionales:

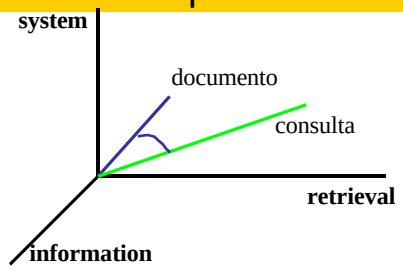
$$d_j = (w_{1j}, w_{2j}, \ldots, w_{tj})$$

Representación de Documentos
Modelo de Espacio de Vectores

Ejemplo:
$D_1 = 2T_1 + 3T_2 + 5T_3$
$D_2 = 3T_1 + 7T_2 + T_3$
$Q = 0T_1 + 0T_2 + 2T_3$



Representación de Documentos
Modelo de Espacio de Vectores



- Los documentos se representan como vectores en un espacio de términos
  - Los términos son usualmente stems
  - Documents represented by binary vectors of terms
- Las consultas se representan en el mismo espacio de documentos
- Se utiliza una medida de distancia de vectores entre la consulta y los documentos

Representación de Documentos
Asignación de Pesos: Binarios

- 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

Representación de Documentos
Asignación de Pesos: Frecuencia

- Los términos más frecuentes en un documento son los más importantes, más indicativos del tema del documento

  f_{ik} = frequency of term i in document k

- Se puede normalizar la frecuencia de un término f_{ik} dividiendola por la frecuencia del término más común en el documento

  $$tf_{ik} = f_{ik} / max_i\{f_{ik}\}$$

Representación de Documentos
Asignación de Pesos: Frecuencia

- 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

Representación de Documentos
Asignación de Pesos: TF x IDF

- TF x 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

Representación de Documentos
Asignación de Pesos: TF x IDF

$$w_{ik} = tf_{ik} * \log(N / n_k)$$

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

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

Representación de Documentos
Asignación de Pesos: TF x IDF

- La Frecuencia Inversa de Documentos (IDF) provee valores altos para palabras raras y bajos para palabras comunes

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

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

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

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

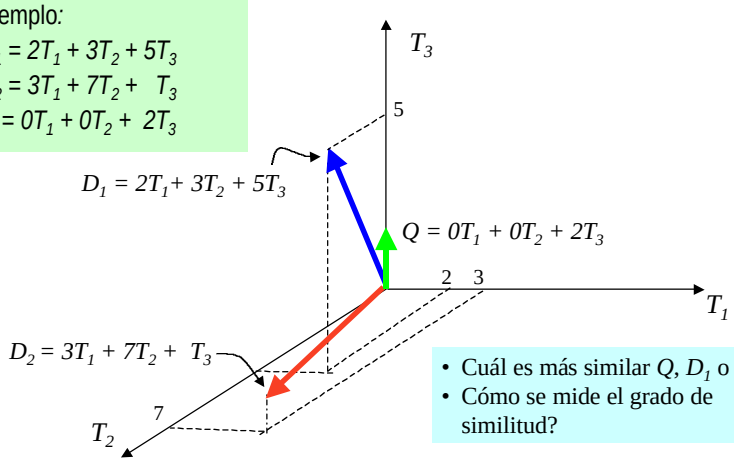
Representación de Documentos
Asignación de Pesos: TF x IDF

- Normalización de los pesos
  - normalizar significa forzar todos los valores para caer dentro de un cierto rango, usualmente entre 0 y 1, inclusive.

$$w_{ik} = \frac{tf_{ik}\log(N/n_k)}{\sqrt{\sum_{k=1}^{t}(tf_{ik})^2[\log(N/n_k)]^2}}$$

Representación de Documentos
Medidas de similitud

Ejemplo:
$D_1 = 2T_1 + 3T_2 + 5T_3$
$D_2 = 3T_1 + 7T_2 + T_3$
$Q = 0T_1 + 0T_2 + 2T_3$



- Cuál es más similar Q, D_1 o D_2?
- Cómo se mide el grado de similitud?

Representación de Documentos
Medidas de similitud

- Una medida de similitud es una función que permiote calcular el grado de similitud de dos vectores.
- Usar una medida de similitud entre una consulta un documento permite:
  - rankear los documentos recuperados en orden de relevancia
  - controlar el número de documentos recuperados mediante el uso de un umbral

Representación de Documentos
Medidas de similitud

$\lvert Q \cap D \rvert$	Matching siple
$2\frac{\lvert Q \cap D \rvert}{\lvert Q \rvert + \lvert D \rvert}$	Coeficiente de Dice
$\frac{\lvert Q \cap D \rvert}{\lvert Q \cup D \rvert}$	Coeficiente de Jaccard
$\frac{\lvert Q \cap D \rvert}{\lvert Q \rvert^{1/2} \times \lvert D \rvert^{1/2}}$	Coseno
$\frac{\lvert Q \cap D \rvert}{\min(\lvert Q \rvert, \lvert D \rvert)}$	Coeficiente de Overlap

Representación de Documentos
Medidas de similitud

$D_i = w_{d_{i1}}, w_{d_{i2}}, ..., w_{d_{it}}$

$Q = w_{q1}, w_{q2,}, ..., w_{qt}$

w = es 0 si un término está ausente

si los pesos están normalizados: $sim(Q, D_i) = \sum_{j=1}^{t} w_{qj} * w_{d_{ij}}$

sino se normaliza durante la comparación de similitud:

$$sim(Q, D_i) = \frac{\sum_{j=1}^{t} w_{qj} * w_{d_{ij}}}{\sqrt{\sum_{j=1}^{t} (w_{qj})^2 * \sum_{j=1}^{t} (w_{d_{ij}})^2}}$$

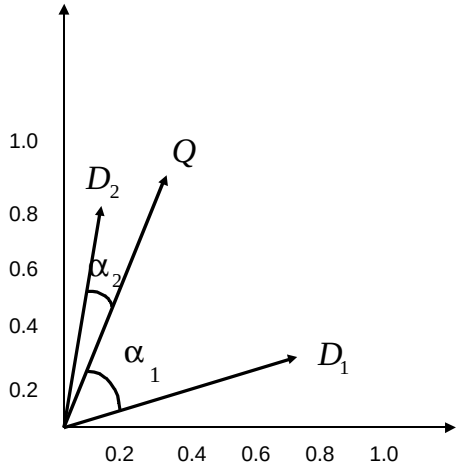
Representación de Documentos
Medidas de similitud

- Dado un vector representando una consulta $Q = (0.4,0.8)$
- y un vector representando un documento $D_2 = (0.2,0.7)$
- se calcula su similitud como:

$$sim(Q, D_2) = \frac{(0.4*0.2)+(0.8*0.7)}{\sqrt{[(0.4)^2+(0.8)^2]*[(0.2)^2+(0.7)^2]}}$$

$$= \frac{0.64}{\sqrt{0.42}} = 0.98$$

Representación de Documentos
Medidas de similitud



$D_1 = (0.8, 0.3)$

$D_2 = (0.2, 0.7)$

$Q = (0.4, 0.8)$

$\cos(\alpha_1) = 0.74$

$\cos(\alpha_2) = 0.98$