Documentos de texto

Los documentos exigen ser representados de una manera estructurada para poder ser preprocesados.

- 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 de texto

- Eliminación de ruido (ej. en documentos HTML) y reducción a un formato ASCII
- Conversión a minúsculas
- Tokenización: divide el texto en entidades significativas (palabras, oraciones, etc.) dados los espacios en blanco presentes y las puntuaciones.
- Eliminación de stop-words
- Stemming
- Identificación de entidades
- Asignación de pesos a términos

Eliminación de ruido

Las tareas de eliminación de ruido en documentos de texto podrían incluir:

- eliminar encabezados de archivos de texto, pies de página
- eliminar HTML, XML, etc. marcado y metadatos
- extraer datos valiosos de otros formatos, como JSON

Tokenización

La tokenización (o segmentación) es un paso que divide cadenas de texto largas en piezas más pequeñas o tokens.

Los trozos de texto más grandes pueden ser convertidos en oraciones, las oraciones pueden ser tokenizadas en palabras, etc.

Entrada: “retrieval, organization and storage”

Salida: tokens

- retrieval
- organization
- and
- storage

Tokenización

Problemas:

- acentuación: *résumé/resume*
- apóstrofes: *L'ensemble*
- abreviaciones: *U.S.A./USA*
- cómo acostumbran los usuarios a escribir las consultas para estas palabras?

Normalización

La normalización generalmente se refiere a una serie de tareas relacionadas destinadas a poner todo el texto en igualdad de condiciones:

- convirtiendo todo el texto en mayúsculas o minúsculas
- eliminando la puntuación
- convirtiendo los números a sus equivalentes de 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.

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 (de su eliminación):

- 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
- reduce la dimensión del espacio de términos

Stemming

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

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

TERMINO	STEM
engineering	engineer
engineered	engineer
engineer	engineer

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"

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

Reconocimiento de entidades

Identificación de elementos en el texto dentro de un conjunto de categorías predefinidas

- Tres categorías aceptadas comúnmente: personas, lugares y organizaciones
- Otros elementos comunes: fechas, medidas (porcentajes, valores monetarios, pero, etc.), direcciones de mail, etc.
- En dominios específicos: nombre de drogas, condiciones médicas, referencias bibliográficas, etc.

Reconocimiento de entidades

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...

Reconocimiento de entidades

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.

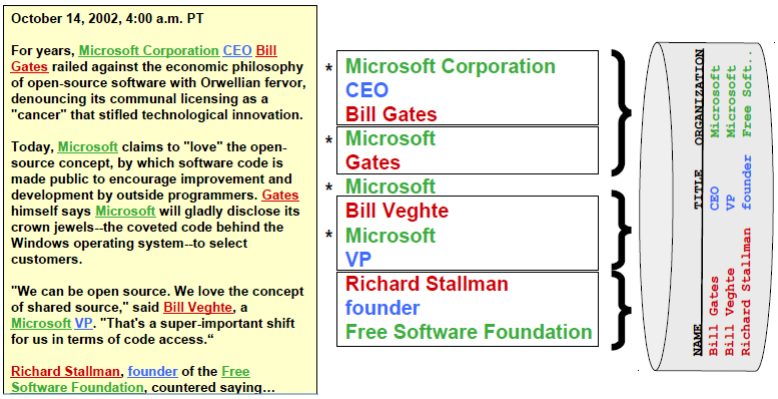
Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"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

Reconocimiento de entidades



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

- 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

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 f_{ik} dividiéndola por la frecuencia del término más común en el documento

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

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

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

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

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

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

diplomatura universitaria en
inteligencia artificial

FACULTAD DE CIENCIAS EXACTAS
UNIVERSIDAD NACIONAL DEL CENTRO DE LA PROVINCIA DE BUENOS AIRES

TP Nº 2 parte 2

Intro a IA: TP 2 parte 2

- Se realizará un práctico utilizando Python según indicaciones en Google Clasroom y clase de consulta por Google Meet