Conceptos Básicos de los Modelos de RI

Índice

- Introducción
- Modelado y Ranking
- Caracterización del modelo de RI
- Taxonomía
- Conceptos básicos
- Ponderación de términos
- Normalización de documentos

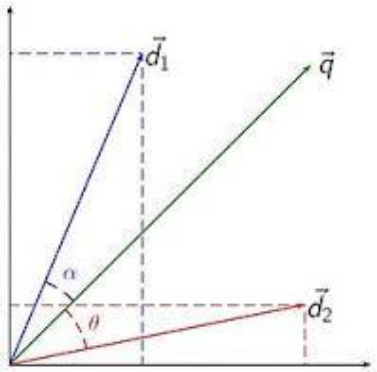
Introducción

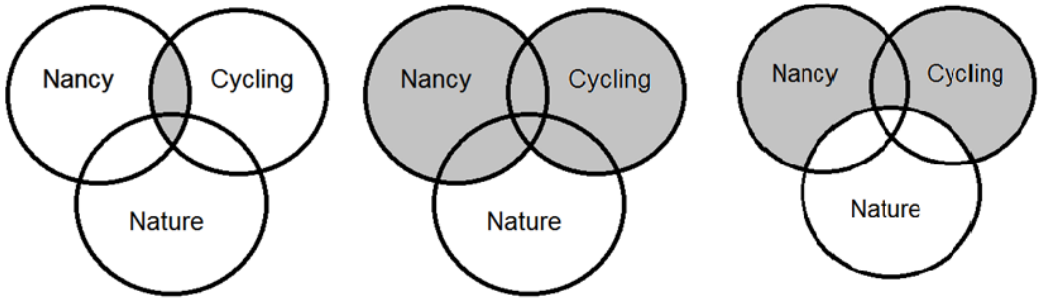
Introducción

- El modelado de sistemas de recuperación es una tarea compleja con el principal **objetivo** de **crear una función de ranking**
  - Función que asigna puntuaciones a los documentos dependiendo de una consulta dada
- El proceso consta de dos tareas principales:
  - El diseño de una estructura lógica para la representación de documentos y consultas
  - La definición de una función de ranking que calcule la posición de cada documento para una consulta dada

Introducción

- La estructura lógica puede basarse en:
  - Conjuntos
  - Vectores
  - Distribución de probabilidades
- El tipo de estructura lógica seleccionada afectará directamente al calculo del ranking.





$$P(t_1 t_2 t_3) = P(t_1) P(t_2 \mid t_1) P(t_3 \mid t_1 t_2)$$

Modelado y Ranking

Modelado y Ranking

- Los sistemas de RI emplean **términos de indexación** (*indexing terms*) para la recuperación e indexación de documentos
  - **Sentido estricto**: palabra clave o grupo de palabras relacionadas con significado propio
    - Generalmente nombre
  - **Sentido general**: cualquier palabra que aparezca en el texto de la colección
- Ventajas:
  - Implementación sencilla
  - Creación de consulta de forma fácil. Reduce el esfuerzo de formulación de la consulta por parte del usuario

Modelado y Ranking

- Desventajas:
  - Reduce la semántica de la consulta
  - Es difícil expresar en pocos términos la necesidad de información de un usuario
- Consecuencias:
  - Los sistemas RI suelen recuperar documentos irrelevantes
  - Los usuarios no saben elegir los términos adecuados para realizar sus consultas
  - Términos erróneos ➔ resultados desastrosos
  - Si los resultados no son relevantes, el usuario no estará satisfecho con el sistema de RI

Modelado y Ranking

- Problema central de la RI:
  - **Predecir** los documentos qué serán **relevantes** para los usuarios
  - **Predecir** los documentos qué serán **irrelevantes** para los usuarios
- Problema con difícil solución:
  - Incertidumbre
  - Vaguedad
  - Los usuarios no tienen porque estar de acuerdo en lo que es o no es relevante
- Los sistemas RI emplean algoritmos de predicción que intentan contentar a la mayor parte de usuarios

Modelado y Ranking

- Algoritmo de predicción:
  - Función de ranking para establecer un orden entre los documentos recuperados
  - Los documentos que se sitúan a la cabeza serán los más relevantes
- Los algoritmos de ranking son la parte central de los sistemas de Recuperación de Información
- Un algoritmo de ranking opera en función de una serie de reglas:
  - Funciones de relevancia
  - Modelo de RI
- Cada combinación dará resultados diferentes

Caracterización del modelo de RI

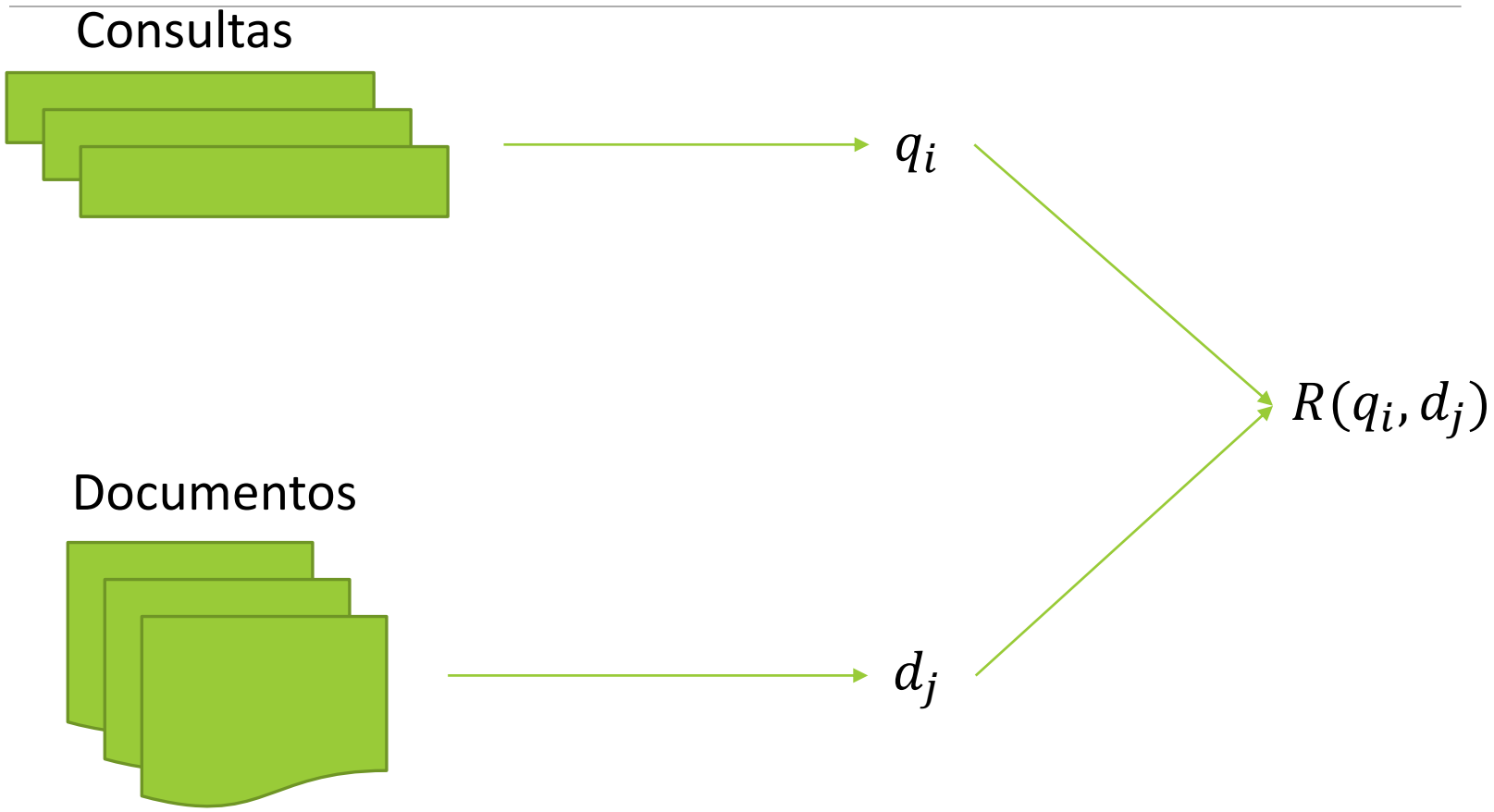
Caracterización del modelo de RI

- **Definición:** un modelo de recuperación de información es un 4-tupla $[\boldsymbol{D}, \boldsymbol{Q}, F, R(q_i, d_j)]$, donde
  - **D** es el conjunto de las vistas o representaciones de los documentos en la colección (corpus)
  - **Q** es el conjunto de las vistas o representaciones de las necesidades de información de los usuarios (consulta)
  - F es el modelo de representación de documentos y consultas, así como sus relaciones (lógicas, vectoriales o probabilísticas)
  - $R(q_i, d_j)$ es una función de ranking que asocia un número real con la representación de una consulta $q_i \in \boldsymbol{Q}$ y la representación de un documento $d_j \in \boldsymbol{D}$. El ranking define un orden entre los documentos en función de la consulta q_i

Caracterización del modelo de RI

- En primer lugar tenemos que definir la representación de los documentos y de las necesidades de información de los usuarios
- Documentos:
  - Subconjunto de los términos del documento
    - Sin palabras vacías
- Consulta:
  - Superconjunto formado por los términos dados por el usuario junto con sinónimos.

Caracterización del modelo de RI



Taxonomía

Taxonomía

- Texto
  - Modelos clásicos (texto no estructurado)
    - Boolean (teoría de conjuntos)
      - Fuzzy
      - Boolean extendido
      - Basado en conjuntos
    - Vector (algebraicos)
      - Modelo vectorial
      - Lantent Semantic Indexing
      - Redes Neuronales
    - Probabilístico
      - BM25
      - Modelos del lenguaje
      - Divergencia aleatoria
      - Redes bayesianas

Taxonomía

- Texto semi-estructurado
  - Proximidad de nodos
  - Basado en XML
- Enlaces (web)
  - Page Rank
  - Hub & Authorities
- Multimedia
  - Imágenes
  - Audio y música
  - Video

Conceptos básicos

Conceptos básicos

TÉRMINOS DE INDEXACIÓN O PALABRAS CLAVE

Conceptos básicos
Términos de indexación

- En los modelos clásicos de IR cada documento se describe como un conjunto de palabras representativas llamadas términos de indexación

- **Definición**: Un término de indexación es *una palabra o grupo de palabras consecutivas en un documento. De forma general, un término de indexación es cualquier palabra en un documento. Este es el enfoque utilizado por los motores de búsqueda. En una interpretación más restrictiva, un término de indexación es un conjunto de palabras preseleccionadas que representan un concepto o tema del documento. Este enfoque es el seguido por los bibliotecarios*.

Conceptos básicos
Términos de indexación

- **Definición**: *Sea t el número de términos de indexación en la colección de documentos y k_i un termino de indexación determinado. $V = \{k_1, \dots k_t\}$ será el conjunto de todos los términos de indexación distintos en de colección y se conocerá por vocabulario V de la colección. El tamaño de la colección es t*
- El vocabulario es una parte muy importante de la colección:
  - Identifica todos los términos de la colección
- Si la colección crece, el vocabulario lo hará también
  - Errores
  - Diferentes formas de expresar un número
  - Acrónimos

Conceptos básicos

DOCUMENTOS Y REPRESENTACIÓN DE CONSULTAS

Conceptos básicos
Documentos y representación de consultas

Definición:

Sea $V = \{k_1, k_2, \dots, k_t\}$ *el vocabulario de una colección.*

Si tres términos de indexación k_l, k_m *y* k_n *aparecen en el mismo documento* d_j, *podemos decir que se ha observado el patrón de coocurrencia* $\{k_l, k_m, k_n\}$.

Para un vocabulario V *de tamaño* t, *el número total de patrones de coocurrencia de términos de la colección es* 2^t.

- *El patrón* $\{1,0, \dots, 0\}$ *indica la presencia del término* k_1, *pero no del resto.*
- *El patrón* $\{1,1, \dots, 1\}$ *indica la presencia de todos los términos.*

Conceptos básicos
Documentos y representación de consultas

- *Cada uno de los patrones de coocurrencia se le llaman componente conjuntivo de términos.*
  - *Un documento d_j, tiene asociado una única componente conjuntiva $c(d_j)$ que describe los términos que aparecen en el documento y cuales no.*
  - *Una consulta q, tiene asociada una única componente conjuntiva $c(q)$ que describe los términos que aparecen en la consulta y cuales no.*
- *La componente conjuntiva $c(d_j)$ es la representación del documento d_j en el sistema.*
- *La componente conjuntiva $c(q)$ es la representación de la consulta q en el sistema.*

Conceptos básicos
Documentos y representación de consultas

- Esta representación de documentos es la más simple que podemos diseñar
  - Bolsa de palabras
- Los modelos de recuperación de información necesitan modelos más complejos y representaciones más sofisticadas de documentos y consultas para mejorar los resultados

Conceptos básicos

MATRIZ DOCUMENTOS-TÉRMINOS

Conceptos básicos
Matriz DxT

- La ocurrencia de un término en un documento establece una relación entre ellos
- La relación término-documento puede cuantificarse:
  - Frecuencia del término en el documento

$$\begin{array}{c} \\ k_1 \\ k_2 \\ k_3 \end{array} \begin{array}{c} \begin{array}{cc} d_1 & d_2 \end{array} \\ \begin{bmatrix} f_{1,1} & f_{1,2} \\ f_{2,1} & f_{2,2} \\ f_{3,1} & f_{3,2} \end{bmatrix} \end{array}$$

Ponderación de términos

Ponderación de términos

- No todos los términos del corpus son igualmente útiles para representar el contenido de los documentos.
- Algunos términos son demasiado “vagos” o genéricos para tenerlos en cuenta.
- Decidir qué términos son adecuados no es una tarea trivial.
- Existen algunas propiedades fácilmente medibles que nos ayudaran a decidir la importancia de un término de indexación.
- Ejemplo: supóngase una colección de 1,000,000 de documentos
  - Una palabra aparece en todos los documentos
    - No es importante ya que la palabra no nos dice nada en particular sobre los documentos
  - Una palabra aparece en sólo 5 documentos
    - La palabra es demasiado específica, sólo está asociada a un porcentaje muy bajo de documentos

Ponderación de términos

- **Solución**: ponderar los términos de los documentos de acuerdo a su importancia.
- **Definición**: *Para caracterizar la importancia de un término, se asocia un peso* $w_{i,j}$, $w_{i,j} > 0$, *a cada término de indexación* k_i, *del documento* d_j *del corpus. Para un término de indexación* k_i *que no aparezca en el documento,* $w_{i,j} = 0$.
- El peso $w_{i,j}$ cuantifica la importancia del término de indexación para describir el contenido de un documento.
- Un sistema de RI tiene que cuantificar la importancia de un término en toda la colección, lo que dependerá de la frecuencia de aparición de los términos en los documentos de la colección.

Ponderación de términos

- **Definición**: *Sea $f_{i,j}$ la frecuencia de aparición del término k_i en el documento d_j, es decir, el número de veces que aparece el término k_i en el documento d_j. La frecuencia global F_i del término k_i en la colección es la suma de las frecuencias del término k_i en todos los documentos*

$$F_i = \sum_{j=1}^{N} f_{i,j}$$

- *Donde, N es el número de documentos de la colección. Al número de documentos en los que aparece el término k_i lo denotaremos como n_i, siendo $n_i \leq F_i$*

Ponderación de términos

CORRELACIÓN ENTRE TÉRMINOS

Ponderación de términos
Correlación entre términos

¿Existe correlación entre los términos de un documento?

¿La frecuencia de aparición es estadísticamente independiente?

Ponderación de términos
Correlación entre términos

- Las frecuencias de los términos de indexación tienen correlación.
  - Las palabras *redes* y *computadores* pueden utilizarse para describir un documento sobre redes de computadores.
  - Están palabras están correlacionadas y sus pesos deberían representar esta correlación.

- La correlación se puede calcular con una matriz de correlación.

Ponderación de términos Correlación entre términos

- Definición. Sea $M = [m_{i,j}]$ una matriz de términos-documentos de t filas por N columnas, donde $m_{i,j} = w_{i,j}$. Es decir, cada celda ij es el peso asociado a la pareja (k_i, d_j). Si M^T es la matriz transpuesta de M, la matriz $C = M \cdot M^T$ será la matriz de correlación término-término. Cada elemento $c_{u,v} \in C$ expresa la correlación entre los términos k_u y k_v.

$$c_{u,v} = \sum_{d_j} w_{u,j} \times w_{v,j}$$

Ponderación de términos
Correlación entre términos

$$\begin{array}{c} \\ k_1 \\ k_2 \\ k_3 \end{array}\begin{array}{c} \begin{array}{cc} d_1 & d_2 \end{array} \\ \begin{bmatrix} w_{1,1} & w_{1,2} \\ w_{2,1} & w_{2,2} \\ w_{3,1} & w_{3,2} \end{bmatrix} \end{array} \qquad \begin{array}{c} \\ d_1 \\ d_2 \end{array}\begin{array}{c} \begin{array}{ccc} k_1 & k_2 & k_3 \end{array} \\ \begin{bmatrix} w_{1,1} & w_{2,1} & w_{3,1} \\ w_{1,2} & w_{2,2} & w_{3,2} \end{bmatrix} \end{array}$$

$$M \qquad \times \qquad M^T$$

$$\begin{array}{c} \\ k_1 \\ k_2 \\ k_3 \end{array}\begin{array}{c} \begin{array}{ccc} k_1 & k_2 & k_3 \end{array} \\ \begin{bmatrix} w_{1,1}w_{1,1} + w_{1,2}w_{1,2} & w_{1,1}w_{2,1} + w_{1,2}w_{2,2} & w_{1,1}w_{3,1} + w_{1,2}w_{3,2} \\ w_{2,1}w_{1,1} + w_{2,2}w_{1,2} & w_{2,1}w_{2,1} + w_{2,2}w_{2,2} & w_{2,1}w_{3,1} + w_{2,2}w_{3,2} \\ w_{3,1}w_{1,1} + w_{3,2}w_{1,2} & w_{3,1}w_{2,1} + w_{3,2}w_{2,2} & w_{3,1}w_{3,1} + w_{3,2}w_{3,2} \end{bmatrix} \end{array}$$

Ponderación de términos
Correlación entre términos

- Asumir una ausencia de correlación entre los términos puede parecer una simplificación demasiado grande
- Sin embargo, la independencia de los términos simplifica enormemente el cálculo de la ponderación, y por ende permite hacer un procesado más rápido
- Actualmente no hay consenso acerca de asumir o no la independencia de términos
- Solución encontrada por la comunidad científica: TF-IDF

Ponderación de términos

TF-IDF

Ponderación de términos TF-IDF

- Esquemas de ponderación más comunes en los sistemas de recuperación de información:
  - Frecuencia de los términos (TF, term frequency)
  - Frecuencia inversa de documentos (IDF, inverse document frequency).
- Ambos esquemas se usan conjuntamente para formar el TF-IDF

Ponderación de términos TF

- **Conjetura de Luhn**: *el valor, o peso, de un término k_i que aparece en un documento d_j es proporcional a su frecuencia $f_{i,j}$. Es decir, cuanto mayor sea el número de veces que la palabra aparece en el documento d_j, mayor será el peso $TF_{i,j}$*

$$tf_{i,j} = f_{i,j}$$

- Variante:

$$tf_{i,j} = \begin{cases} 1 + \log f_{i,j} & Si f_{i,j} > 0 \\ 0 & en\ otro\ caso \end{cases}$$

- Usaremos la variante basada en la expresión logarítmica como forma predeterminada para el cálculo del TF.

Ponderación de términos TF

- Supongamos el siguiente corpus:

To do is to be. To be is to do.	To be or not to be. I am what I am.	I think therefore I am. Do be do be do.	Do do do, da da da. Let it be, let it be.
d_1	d_2	d_3	d_4

Ponderación de términos TF

#	Término	$f_{i,1}$	$f_{i,2}$	$f_{i,3}$	$f_{i,4}$	$tf_{i,1}$	$tf_{i,2}$	$tf_{i,3}$	$tf_{i,4}$
1	to	4	2			3	2		
2	do	2		3	3	2		2,585	2,585
3	is	2				2			
4	be	2	2	2	2	2	2	2	2
5	or		1				1		
6	not		1				1		
7	I		2	2			2	2	
8	am		2	1			2	1	
9	what		1				1		
10	think			1				1	
11	therefore			1				1	
12	da				3				2,585
13	let				2				2
14	it				2				2
Tamaño documentos		10	11	10	12				

Ponderación de términos IDF

- *La **exhaustividad** (exhaustivity) es una propiedad de la descripción de un documento. Debe interpretarse como la importancia que tiene dicha descripción para el tema general del documento.*
- *La **particularidad** (specificity) es una propiedad de los términos de indexación. Debe interpretarse como la bondad de los términos para describir el tema del documento.*
- Si añadimos nuevos términos al documento, la exhaustividad de la descripción del documento se incrementará.
  - La probabilidad de que el documento se corresponda con la consulta también se incrementa

Ponderación de términos IDF

- **Definición**. ***Exhaustividad Óptima****: cuantos más términos se le asignen a un documento, su descripción será más exhaustiva. La probabilidad de ser recuperado por una consulta aleatoria, también será mayor. Sin embargo, si a un documento se le asignan muchos términos, será recuperado para muchas consultas para las que el documento no es relevante. Esto nos sugiere que el número medio de términos de indexación por documento debería de optimizarse para que la probabilidad de ser relevante sea maximizada. El número optimo de términos de indexación define la exhaustividad óptima para la descripciones de los documentos.*

Ponderación de términos IDF

- **Definición**. ***Exhaustividad y particularidad estadística****: en términos estadísticos, la exhaustividad de la descripción de un documento puede cuantificarse como el número de términos de indexación que contiene el documento. Consecuentemente, la particularidad de un término puede cuantificarse como la función inversa del número de documentos en los que aparece.*
- Relación entre exhaustividad y particularidad
  - Si la descripción de un documento crece, la particularidad de los términos decrece.
  - Si un término aparece en todos los documentos, su particularidad es mínima, por lo que el término no es útil para la recuperación
- La ponderación de términos puede representarse como una función de la particularidad de los términos.

Ponderación de términos IDF

- La frecuencia relativa de las palabras de un documento puede aproximarse mediante una distribución matemática.
- ***Ley de Zipf**: Sea n_i la frecuencia del término k_i en el corpus. Ordenemos las frecuencias de los términos en orden decreciente. Sea $n(r)$ el r-ésimo término con mayor frecuencia. De este modo, de acuerdo con la Ley de Zipf $n(r) \sim r^{-\alpha}$, donde α es una constante determinada empíricamente. Es decir, la frecuencia de un término en la colección puede modelarse mediante una función exponencial de su rango.*
- La Ley de Zipf puede expresarse de la siguiente forma:

$$n(r) = C \cdot r^{-\alpha}$$

- Donde C es una segunda constante que se determinará de forma empírica.

Ponderación de términos IDF

- Para el idioma inglés, α puede aproximarse a 1

$$\log n(r) = \log C + \log r$$

- Para $r = 1$, se obtiene

$$C = n(1)$$

- Donde C es la mayor frecuencia de la colección y actúa como una constante de normalización.
- Una simplificación sería normalizar utilizando la mayor frecuencia posible. Es decir, $C = N$, donde N es el número total de documentos en la colección.

$$\log r \sim \log N - \log n(r)$$

Ponderación de términos IDF

- Dicha ley de puede emplearse para ponderar los términos de la colección.
- *Sea k_i el término con la r-ésima mayor frecuencia. Es decir, $n(r) = n_i$. El peso IDF se puede asociar de la forma:*

$$IDF_i = log \frac{N}{n_i}$$

- *Donde IDF_i se conoce como frecuencia inversa del término.*

Ponderación de términos IDF

#	Término	n_i	$IDF = log\ (\mathrm{N}/n_i)$
1	to	2	1
2	do	3	0,415
3	is	1	2
4	be	4	0
5	or	1	2
6	not	1	2
7	I	2	1
8	am	2	1
9	what	1	2
10	think	1	2
11	therefore	1	2
12	da	1	2
13	let	1	2
14	it	1	2

Ponderación de términos TF-IDF

- Los sistemas de ponderación mas populares usan una combinación de la frecuencia inversa y la frecuencia dentro del documento.
- Definición. Sea $w_{i,j}$ el peso del término asociado a la pareja (k_i, d_j). Entonces, la ponderación TF-IDF puede definirse como

$$w_{i,j} = \begin{cases} (1 + \log f_{i,j}) \times \log \frac{N}{n_i} & Si\ f_{i,j} > 0 \\ 0 & en\ otro\ caso \end{cases}$$

Ponderación de términos TF-IDF

#	Término	$tf_{i,1}$	$tf_{i,2}$	$tf_{i,3}$	$tf_{i,4}$	IDF	d_1	d_2	d_3	d_4
1	to	3	2			1	3	2		
2	do	2		2,585	2,585	0,415	0,830		1,073	1,073
3	is	2				2	4			
4	be	2	2	2	2	0				
5	or		1			2		2		
6	not		1			2		2		
7	I		2	2		1		2	2	
8	am		2	1		1		2	1	
9	what		1			2		2		
10	think			1		2			2	
11	therefore			1		2			2	
12	da				2,585	2				5,170
13	let				2	2				4
14	it				2	2				4

Ponderación de términos TF-IDF

- Variantes de TF:
  - Binaria: $\{0, 1\}$
  - Frecuencia bruta: $\{f_{i,j}\}$
  - Normalización logarítmica: $1 + \log f_{i,j}$
  - Doble normalización 0,5: $0{,}5 + 0{,}5 \frac{f_{i,j}}{max_i f_{i,j}}$
  - Doble normalización K: $\mathrm{k} + (1 - K) \frac{f_{i,j}}{max_i f_{i,j}}$

Ponderación de términos TF-IDF

- Variantes de IDF:
  - Unitaria: $\{0, 1\}$
  - Frecuencia inversa: $\log \frac{N}{n_i}$
  - Frecuencia inversa suavizada: $\log\left(1 + \frac{N}{n_i}\right)$
  - Frecuencia inversa máxima: $\log\left(1 + \frac{max_i\, n_i}{n_i}\right)$
  - Frecuencia inversa probabilística: $\log \frac{N - n_i}{n_i}$

Ponderación de términos TF-IDF

- Variantes de $TF - IDF$:
  - Esquema 1:
    - Términos documentos: $f_{i,j} \times \log \frac{N}{n_i}$
    - Términos consulta: $(0{,}5 + 0{,}5 \frac{f_{i,q}}{max_i f_{i,q}})$
  - Esquema 2:
    - Términos documentos: $1 + \log f_{i,j}$
    - Términos consulta: $\log\left(1 + \frac{N}{n_i}\right)$
  - Esquema 3:
    - Términos documentos: $(1 + \log f_{i,j}) \times \log \frac{N}{n_i}$
    - Términos consulta: $(1 + \log f_{i,q}) \times \log \frac{N}{n_i}$

Normalización de documentos

Normalización de documentos

- En colección grandes puede haber una gran diferencia en el tamaño de los documentos.
  - Los documentos grandes tiende a ser más recuperados por el sistema.
- Solución:
  - Normalizar el ranking calculado por el sistema en función de la longitud del documento.
- **Definición**. ***Tamaño en bytes***. *Considérese que el documento se representa simplemente como un conjunto de bytes. En este caso, la longitud del documento será el número de bytes de ese conjunto, es decir, el tamaño del documento. La principal ventaja de este enfoque es su simplicidad.*

Normalización de documentos

- **Definición**. ***Tamaño en palabras***. *Considérese que el documento se representa como una cadena de caracteres que puede dividirse en palabras. En este caso, la longitud del documento será el número de palabras que lo compone. Este representación es simple, pero calcula la longitud del documento a nivel sintáctico, lo que añade una mayor semántica.*

Normalización de documentos

- **Definición**. ***Vector normal***. *Considérese que cada término es asociado a un con el vector unitario ortonormal k_i en un espacio t-dimensional (donde t es el número total de términos). En este espacio, los documentos pueden representarse como vectores de términos ponderados. El término k_i del documento d_j se asocia al vector $w_{i,j} \times \overrightarrow{k_i}$ y representa la contribución del término al documento. La representación del documento d_j será el vector formado por todos los términos (componentes vectoriales). La longitud del documento se calculará de la siguiente forma:*

$$|\overrightarrow{d_j}| = \sqrt{\sum_{i}^{t} w_{i,j}^2}$$

Normalización de documentos

Longitud	d_1	d_2	d_3	d_4
Bytes	34	37	41	43
Palabras	10	11	10	12
Vector normal	5,068	4,899	3,762	7,738