FACULTAD DE CIENCIAS DE LA DOCUMENTACIÓN

GRADO EN INFORMACIÓN Y DOCUMENTACIÓN

MODELOS AVANZADOS DE RECUPERACIÓN

DE LA INFORMACIÓN

CUADERNO DE TRABAJO

Nº17

Profesor

Juan Antonio Martínez Comeche

Colección “Cuadernos de Trabajo”, n º 17

Grado en Información y Documentación

Coordinador de la Titulación: Fátima Martín Escudero

Coordinador de la Colección: José Luis Gonzalo Sánchez-Molero

Septiembre 2016

ISBN-13: 978-84-608-9061-4

Depósito Legal: M-33848-2016

Edita: Facultad de Ciencias de la Documentación

Universidad Complutense

C/ Santísima Trinidad, 37

28010 Madrid

Índice

INTRODUCCIÓN

Este nuevo cuaderno de trabajo es la continuación del Cuaderno de Trabajo nº 10 titulado "Modelos clásicos de Recuperación de la Información", publicado en 2013. Si entonces el principal objetivo era abordar los modelos clásicos de Recuperación de Información (RI), exponiendo los principios teóricos en que se basan junto con algunos ejercicios que permitiesen su asimilación, en esta ocasión nos proponemos exponer los inconvenientes más destacados que presentan dichos modelos clásicos y las principales modificaciones técnicas que se han propuesto para superar dichos inconvenientes.

De este modo, ponemos a disposición de los alumnos de la asignatura "Búsqueda y Recuperación de Información", asignatura obligatoria del Grado en Información y Documentación, un cuaderno en el que podrá consultar los principios teóricos esenciales que rigen el funcionamiento actual de los Sistemas de Recuperación de Información, al tiempo que se explica el grado de innovación que implican en relación a los primeros modelos de Recuperación de Información desarrollados, los denominados modelos clásicos. Esta primera etapa en la superación de los modelos clásicos, de donde el nombre de modelos avanzados de Recuperación de Información, constituyen actualmente una referencia sobre la que comparar nuevas técnicas de Recuperación de Información. En especial el modelo BM25 se ha venido considerando un hito importante sobre el que medir el avance de nuevas técnicas experimentales en esta área de conocimiento. Como veremos a lo largo del texto, la versión correspondiente del modelo vectorial presenta un alto grado de semejanza con el modelo BM25, reforzando la idea de unos principios teóricos comunes compartidos en estas primeras aproximaciones desarrolladas a raíz de los modelos clásicos.

Con el ánimo de que el alumno observe la evolución desde cada uno de los modelos clásicos, y al tiempo comprenda cómo finalmente confluyen en soluciones semejantes, este cuaderno de trabajo se divide en tres secciones que corresponden a cada uno de los modelos clásicos:

- Modelo booleano extendido
- Modelo BM25/Okapi, extensión del modelo probabilístico clásico
- Modelo vectorial con normalización por longitud basada en pivote

En cada una de estas secciones se explica muy brevemente las características técnicas y los principios teóricos en los que se basa el modelo correspondiente, insistiendo en los inconvenientes del modelo clásico precedente que trata de superar.

A su vez, tras la exposición teórica, cada modelo viene acompañado de una serie de ejercicios posteriores en número suficiente para alcanzar una comprensión del modelo correspondiente.

Los ejercicios incluyen, en relación a cada modelo, diversos modos de representación de los datos de carácter estadístico esenciales en el tratamiento automatizado de la Recuperación de información textual, completando de este modo el ciclo completo que va desde la representación hasta la recuperación de documentos textuales. Esperamos que, de este modo, se facilite la comprensión del proceso íntegro a que se somete una colección textual de cara a la recuperación automatizada de la información mediante la introducción de consultas por parte de los usuarios.

MODELO BOOLEANO EXTENDIDO

El modelo booleano extendido trata de superar la principal limitación del modelo booleano clásico, esto es, la imposibilidad de ordenar los documentos de la respuesta en relación a una consulta.

El modelo booleano clásico muestra los documentos que satisfacen la fórmula booleana de la consulta, eliminando de la respuesta el resto de los documentos de la colección. A su vez, el modelo booleano clásico es incapaz de inferir, de entre los documentos que satisfacen la consulta, cuál es el orden en que deben mostrarse al usuario, considerando su nivel de relevancia o similaridad con la consulta.

El modelo booleano extendido trata de superar esta limitación imponiendo pesos a los términos en los documentos; es decir, imponiendo un número positivo a cada término en cada documento, número directamente proporcional a su importancia para representar el contenido del documento. Por ejemplo, si en un cierto documento D un término A tiene peso 0'26 y otro término B tiene peso 0'13, implica que -conforme al procedimiento empleado para imponer estos pesos- el término A representa en mayor medida el contenido del documento D y que en menor medida el documento D versa sobre el término B. Siguiendo con este ejemplo, si en otro documento F el término A tiene peso 0'32 y el término B tiene peso 0'28, el sistema podrá responder a la consulta Q = A OR B no solo indicando que los documentos D y F satisfacen la consulta, sino que el documento F (con pesos 0'32 y 0'28) satisface en mayor medida dicha consulta que el documento D (con pesos 0'26 y 0'13), pudiendo de esta manera ordenar los documentos en la respuesta.

Esta conclusión lógica se explica matemáticamente si consideramos un eje de coordenadas donde el eje X representa uno de los términos (el término A, por ejemplo) y el eje Y representa el otro término (el término B, en este caso). Para mantener las nociones básicas del modelo booleano clásico, el eje X (correspondiente al término A) puede variar entre 0 y 1, al igual que el eje Y (correspondiente al término B). Conforme a este esquema, cada documento está representado como un punto en dicho espacio. Así, el documento D estaría representado por el punto D(0'26, 0'13), mientras que el documento F estaría representado por el punto F(0'32, 0'28).

Siguiendo en este plano, convendremos en que ante una consulta Q = A OR B, un documento situado en el punto origen O(0,0) sería totalmente irrelevante (su contenido no versa en absoluto ni sobre el término A ni sobre el término B), mientras que otro documento situado en el punto (1,1) sería totalmente relevante (pues su contenido versa plenamente tanto sobre el término A como sobre el término B). Además, cuanto más alejado del origen (0,0), en mayor medida el documento verifica que su contenido versa sobre A o sobre B y, por tanto, más cumple la consulta en OR dicho documento. Esto sugirió a los desarrolladores que, para una consulta tipo OR, se tomase la distancia euclidiana (la que se mide considerando una línea recta entre dos puntos) entre el punto origen O(0,0) y el punto correspondiente al documento como medida de la relevancia/similaridad de dicho documento con respecto a la consulta. De esta manera, cuanto mayor es la distancia entre el documento y el origen O(0,0), mayor es la relevancia/similaridad de dicho documento en relación a una consulta tipo OR.

Ante una consulta tipo AND, curiosamente, un documento totalmente relevante estaría igualmente situado en el punto (1,1), pues versa sobre ambos términos simultáneamente y en la mayor medida posible. Del mismo modo, un documento situado en el origen O(0,0) sería totalmente irrelevante, pues no versa sobre ninguno de los términos de la consulta. La diferencia con la consulta tipo OR radica en que ahora, en una consulta tipo AND, cuanto más alejado esté el documento del punto (1,1), en menor medida el documento verifica que su contenido versa simultáneamente sobre A y B. En consecuencia, en una consulta tipo AND cuanto menor es la distancia entre el documento y el punto (1,1) mayor es la relevancia/similaridad de dicho documento en relación a dicha consulta.

En resumen, en el modelo booleano extendido las similaridades entre documento y consulta se calculan de manera que:

- Ante una consulta tipo Q = A OR B, cuanto mayor es la distancia entre el documento y el origen O(0,0), mayor es la relevancia/similaridad de dicho documento en relación a la consulta.
- Ante una consulta tipo Q = C AND F, cuanto menor es la distancia entre el documento y el punto (1,1), mayor es la relevancia/similaridad de dicho documento en relación a la consulta.

Empleando la distancia euclidiana normalizada, tenemos las siguientes fórmulas:

$$SIM(OR)(q,d) = \sqrt{\frac{d_1^2 + d_2^2}{2}}$$

$$SIM(AND)(q,d) = 1 - \sqrt{\frac{(1-d_1)^2 + (1-d_2)^2}{2}}$$

Teniendo en cuenta que:

- En la colección existen ‘t’ términos distintos.
- Cada documento posee ‘t’ pesos, uno para cada término de la colección.
- ‘d_1’ y ‘d_2’ son los pesos de los dos términos de la consulta ‘q’ en el documento ‘d’.

BIBLIOGRAFÍA

Baeza-Yates, R.; Ribeiro-Neto, B. (2011). Extended boolean model. En: Modern information retrieval: the concepts and technology behind search, Harlow: Pearson, 2011, pp. 92-95.

Fox, E.A. (1983). Extending the boolean and vector space models of information retrieval with P-Norm queries and multiple concept types. PhD thesis, Cornell University.

Salton, G.; Fox, E.A.; Wu, H. (1983). Extended boolean information retrieval. Communications of the ACM, 26(11): 1022-1036.

EJERCICIO 1

Sea un Sistema de Recuperación de información cuya colección consta de un millón de documentos sobre Biblioteconomía. La última consulta realizada al sistema incluía los términos "CDU", "Dewey" y "auxiliares". Los datos estadísticos esenciales en relación a dichos términos se resumen en la siguiente tabla:

	n	Lista
CDU	18709	(49,3),(67,1),(68,4),(90,2)
Dewey	12006	(67,4),(78,1),(99,2)
auxiliares	25413	(49,6),(68,4),(81,7),(83,5),(90,2)

Donde 'n' es el número de documentos de la colección en los que aparece un determinado término. En la 'Lista' se incluyen exclusivamente los números de los 100 primeros documentos de la colección que contienen cada término, así como la frecuencia de aparición del término en cada documento. Así, (79,3) indica que el término en cuestión aparece 3 veces en el documento número 79.

El SRI se basa en el modelo booleano extendido y el peso de los términos en los documentos se calcula como el cociente entre la frecuencia del término en el documento y la frecuencia máxima de un término en un documento cualquiera de la colección (en este caso, considere 7 como dicha frecuencia máxima).

Es suficiente utilizar dos decimales en los cálculos. A la hora de representar la respuesta, se indicará con el símbolo '/' el orden entre documentos, y una coma servirá para indicar que el sistema no puede imponer un orden entre documentos (debido a que poseen la misma similaridad con la consulta). Así, la respuesta D3 / D5 / D2, D4 / D1 indicaría que el sistema situaría en primer lugar, con la máxima similaridad, el documento D3; a continuación, con menor similaridad, el documento D5; a continuación, con la misma similaridad –aunque menor que la que posee el documento D5-, el sistema incluye los documentos D2 y D4; por último, con la menor similaridad, el documento D1.

Hallar la respuesta del sistema en relación a los 100 primeros documentos de la colección ante las consultas:

Q1 = CDU AND auxiliares

Q2 = CDU OR Dewey

SOLUCIÓN

Dividiendo cada una de las frecuencias por 7 obtenemos los pesos de cada término en los 100 primeros documentos. Dicha información la resumimos en la siguiente tabla:

	N	LISTA
CDU	18709	(49,0'43),(67,0'14),(68,0'57),(90,0'29)
Dewey	12006	(67,0'57),(78,0'14),(99,0'29)
auxiliares	25413	(49,0'86),(68,0'57),(81,1),(83,0'71),(90,0'29)

Q1 = CDU AND auxiliares

Calculamos la similaridad correspondiente a la conectiva AND entre la consulta Q1 y cada uno de los 100 primeros documentos de la colección, considerando en el eje X los pesos de los documentos en relación al término 'CDU' y en el eje Y los pesos de los documentos en relación al término 'auxiliares':

SIM (AND) (Q1, D49) = $1 - \sqrt{\frac{(1-D49_{CDU})^2+(1-D49_{auxiliares})^2}{2}} = 1 - \sqrt{\frac{(1-0,43)^2+(1-0,86)^2}{2}} = 0'59$

SIM (AND) (Q1, D67) = $1 - \sqrt{\frac{(1-D67_{CDU})^2+(1-D67_{auxiliares})^2}{2}} = 1 - \sqrt{\frac{(1-0,14)^2+(1-0)^2}{2}} = 0'07$

SIM (AND) (Q1, D68) = $1 - \sqrt{\frac{(1-D68_{CDU})^2+(1-D68_{auxiliares})^2}{2}} = 1 - \sqrt{\frac{(1-0,57)^2+(1-0,57)^2}{2}}$ = 0'57

SIM (AND) (Q1, D81) = $1 - \sqrt{\frac{(1-D81_{CDU})^2+(1-D81_{auxiliares})^2}{2}} = 1 - \sqrt{\frac{(1-0)^2+(1-1)^2}{2}}$ = 0'29

SIM (AND) (Q1, D83) = $1 - \sqrt{\frac{(1-D83_{CDU})^2+(1-D83_{auxiliares})^2}{2}} = 1 - \sqrt{\frac{(1-0)^2+(1-0,71)^2}{2}}$ = 0'27

SIM (AND) (Q1, D90) = $1 - \sqrt{\frac{(1-D90_{CDU})^2+(1-D90_{auxiliares})^2}{2}} = 1 - \sqrt{\frac{(1-0,29)^2+(1-0,29)^2}{2}}$ = 0'29

El resto de los 100 primeros documentos de la colección tienen un valor de similaridad 0 con la consulta Q1, por lo que los eliminamos de la respuesta.

Por tanto, la respuesta del Sistema a Q1 sería la siguiente: D49 / D68 / D81, D90 / D83 / D67

Q2 = CDU OR Dewey

Calculamos la similaridad correspondiente a la conectiva OR entre la consulta Q2 y cada uno de los 100 primeros documentos de la colección, considerando en el eje X los pesos de los documentos en relación al término 'CDU' y en el eje Y los pesos de los documentos en relación al término 'Dewey':

SIM (OR) (Q2, D49) = $\sqrt{\frac{D49_{CDU}^2 + D49_{Dewey}^2}{2}} = \sqrt{\frac{0{,}43^2 + 0^2}{2}} = 0'30$

SIM (OR) (Q2, D67) = $\sqrt{\frac{D67_{CDU}^2 + D67_{Dewey}^2}{2}} = \sqrt{\frac{0{,}14^2 + 0{,}57^2}{2}} = 0'41$

SIM (OR) (Q2, D68) = $\sqrt{\frac{D68_{CDU}^2 + D68_{Dewey}^2}{2}} = \sqrt{\frac{0{,}57^2 + 0^2}{2}} = 0'40$

SIM (OR) (Q2, D78) = $\sqrt{\frac{D78_{CDU}^2 + D78_{Dewey}^2}{2}} = \sqrt{\frac{0^2 + 0{,}14^2}{2}} = 0'10$

SIM (OR) (Q2, D90) = $\sqrt{\frac{D90_{CDU}^2 + D90_{Dewey}^2}{2}} = \sqrt{\frac{0{,}29^2 + 0^2}{2}} = 0'21$

SIM (OR) (Q2, D99) = $\sqrt{\frac{D99_{CDU}^2 + D99_{Dewey}^2}{2}} = \sqrt{\frac{0^2 + 0{,}29^2}{2}} = 0'21$

El resto de los 100 primeros documentos de la colección tienen un valor de similaridad 0 con la consulta Q2, por lo que los eliminamos de la respuesta.

Por tanto, la respuesta del Sistema a Q2 sería la siguiente: D67 / D68 / D49 / D90, D99 / D78

En resumen, la respuesta del Sistema a Q1 sería: D49 / D68 / D81, D90 / D83 / D67

Y la respuesta del Sistema a Q2 sería: D67 / D68 / D49 / D90, D99 / D78

EJERCICIO 2

Sea un Sistema de Recuperación de Información cuya colección consta de un millón de documentos sobre Documentación automatizada. Los datos estadísticos esenciales en relación a los términos de indexación empleados en las consultas se resumen en la siguiente tabla:

	N	LISTA DOCUMENTOS (HASTA NÚMERO 1000)	LISTA APARICIONES (NÚMDOC, FREC, <POSIC>)
t1	47631	28, 140, 665, 789, 804	(28,1,<28>), (140,2,<5,37>), (665,1,<241>), (789,3,<4,16,58>), (804, 2,<327,459>)
t2	148725	610, 789, 864	(610,2,<69,87>), (789,1,<221>), (864,1,<191>)
t3	3499	25, 335, 368, 879, 902	(25,3,<44,72,120>), (335,2,<4,8>), (368,1,<499>), (879,1,<703>), (902,2,<59,66>)
t4	187	100, 140, 206, 442	(100,1,<208>), (140,1,<80>), (206,2,<21,108>), (442,1,<273>)
t5	36210	114, 123, 335, 789	(114,1,<31>), (123,2,<75,81>), (335,4,<5,9,26,32>), (789,1,<102>)
t6	577024	16, 98, 206	(16,1,<667>), (98,2,<238,304>), (206,1,<717>)

Donde 'n' es el número de documentos de la colección en los que aparece un determinado término. En la 'Lista documentos' se incluyen exclusivamente los números de los 1000 primeros documentos de la colección que contienen cada término. En la 'Lista apariciones' se incluye el número de documento, la frecuencia de aparición del término en dicho documento y la/s posición/es absoluta/s de aparición desde el inicio de cada documento. Así, (79,3, <156, 208,367>) indica que el término en cuestión aparece 3 veces en el documento número 79, en las posiciones 156, 208 y 367 contando las palabras desde el inicio del documento.

El SRI se basa en el modelo booleano extendido y el peso de los términos en los documentos se calcula como el cociente entre la frecuencia del término en el documento y la frecuencia máxima de un término en un documento cualquiera de la colección (en este caso, considere 4 como dicha frecuencia máxima).

Es suficiente utilizar dos decimales en los cálculos. A la hora de representar la respuesta, se indicará con el símbolo '/' el orden entre documentos, y una coma servirá para indicar que el sistema no puede imponer un orden entre documentos (debido a que poseen la misma similaridad con la consulta). Así, la respuesta D3 / D5 / D2, D4 / D1 indicaría que el sistema situaría en primer lugar, con la máxima similaridad, el documento D3; a continuación, con menor similaridad, el documento D5; a continuación, con la misma similaridad –aunque menor que la que posee el documento D5-, el sistema incluye los documentos D2 y D4; por último, con la menor similaridad, el documento D1.

Hallar la respuesta del sistema en relación a los 1000 primeros documentos de la colección ante las consultas Q1 = (t1 AND t5) y Q2 = (t4 OR t6)

SOLUCIÓN

Dividiendo cada una de las frecuencias por 4 obtenemos los pesos de cada término en los 1000 primeros documentos de la colección. Como en las consultas del ejercicio no se emplean los términos t2 y t3, la información relativa a dichos términos se elimina en la tabla resultante:

	n	Lista documentos (hasta número 1000)	Lista apariciones (númDoc, frec, <posic>)
t1	47631	28, 140, 665, 789, 804	(28,0'25,<28>), (140,0'5,<5,37>), (665,0'25,<241>), (789,0'75,<4,16,58>), (804, 0'5,<327,459>)
t4	187	100, 140, 206, 442	(100,0'25,<208>), (140,0'25,<80>), (206,0'5,<21,108>), (442,0'25,<273>)
t5	36210	114, 123, 335, 789	(114,0'25,<31>), (123,0'5,<75,81>), (335,1,<5,9,26,32>), (789,0'25,<102>)
t6	577024	16, 98, 206	(16,0'25,<667>), (98,0'5,<238,304>), (206,0'25,<717>)

Q1 = t1 AND t5

Calculamos la similaridad correspondiente a la conectiva AND entre la consulta Q1 y cada uno de los 1000 primeros documentos de la colección, considerando en el eje X los pesos de los documentos en relación al término t1 y en el eje Y los pesos de los documentos en relación al término t5:

$SIM_{AND}(Q1, D28) = 1 - \sqrt{\frac{(1-D28_{t1})^2+(1-D28_{t5})^2}{2}} = 1 - \sqrt{\frac{(1-0,25)^2+(1-0)^2}{2}} = 0'12$

$SIM_{AND}(Q1, D114) = 1 - \sqrt{\frac{(1-D114_{t1})^2+(1-D114_{t5})^2}{2}} = 1 - \sqrt{\frac{(1-0)^2+(1-0,25)^2}{2}} = 0'12$

$SIM_{AND}(Q1, D123) = 1 - \sqrt{\frac{(1-D123_{t1})^2+(1-D123_{t5})^2}{2}} = 1 - \sqrt{\frac{(1-0)^2+(1-0,5)^2}{2}} = 0'21$

$SIM_{AND}(Q1, D140) = 1 - \sqrt{\frac{(1-D140_{t1})^2+(1-D140_{t5})^2}{2}} = 1 - \sqrt{\frac{(1-0,5)^2+(1-0)^2}{2}} = 0'21$

$SIM_{AND}(Q1, D335) = 1 - \sqrt{\frac{(1-D335_{t1})^2+(1-D335_{t5})^2}{2}} = 1 - \sqrt{\frac{(1-0)^2+(1-1)^2}{2}} = 0'29$

$SIM_{AND}(Q1, D665) = 1 - \sqrt{\frac{(1-D665_{t1})^2+(1-D665_{t5})^2}{2}} = 1 - \sqrt{\frac{(1-0,25)^2+(1-0)^2}{2}} = 0'12$

$SIM_{AND}(Q1, D789) = 1 - \sqrt{\frac{(1-D789_{t1})^2+(1-D789_{t5})^2}{2}} = 1 - \sqrt{\frac{(1-0,75)^2+(1-0,25)^2}{2}} = 0'44$

$SIM_{AND}(Q1, D804) = 1 - \sqrt{\frac{(1-D804_{t1})^2+(1-D804_{t5})^2}{2}} = 1 - \sqrt{\frac{(1-0,5)^2+(1-0)^2}{2}} = 0'21$

El resto de los 1000 primeros documentos de la colección tienen un valor de similaridad 0 con la consulta Q1, por lo que los eliminamos de la respuesta.

Por tanto, la respuesta del Sistema a Q1 sería la siguiente:

D789 / D335 / D123, D140, D804 / D28, D114, D665

Q2 = t4 OR t6

Calculamos la similaridad correspondiente a la conectiva OR entre la consulta Q2 y cada uno de los 1000 primeros documentos de la colección, considerando en el eje X los pesos de los documentos en relación al término t4 y en el eje Y los pesos de los documentos en relación al término t6:

$SIM_{OR}(Q2, D16) = \sqrt{\frac{D16_{t4}^2 + D16_{t6}^2}{2}} = \sqrt{\frac{0^2 + 0{,}25^2}{2}} = 0'18$

$SIM_{OR}(Q2, D98) = \sqrt{\frac{D98_{t4}^2 + D98_{t6}^2}{2}} = \sqrt{\frac{0^2 + 0{,}5^2}{2}} = 0'35$

$SIM_{OR}(Q2, D100) = \sqrt{\frac{D100_{t4}^2 + D100_{t6}^2}{2}} = \sqrt{\frac{0{,}25^2 + 0^2}{2}} = 0'18$

$SIM_{OR}(Q2, D140) = \sqrt{\frac{D140_{t4}^2 + D140_{t6}^2}{2}} = \sqrt{\frac{0{,}25^2 + 0^2}{2}} = 0'18$

$SIM_{OR}(Q2, D206) = \sqrt{\frac{D206_{t4}^2 + D206_{t6}^2}{2}} = \sqrt{\frac{0{,}5^2 + 0{,}25^2}{2}} = 0'40$

$SIM_{OR}(Q2, D442) = \sqrt{\frac{D442_{t4}^2 + D442_{t6}^2}{2}} = \sqrt{\frac{0{,}25^2 + 0^2}{2}} = 0'18$

El resto de los 1000 primeros documentos de la colección tienen un valor de similaridad 0 con la consulta Q2, por lo que los eliminamos de la respuesta.

Por tanto, la respuesta del Sistema a Q2 sería la siguiente:

D206 / D98 / D16, D100, D140, D442

En resumen, la respuesta a Q1 sería: D789 / D335 / D123, D140, D804 / D28, D114, D665

Y la respuesta del Sistema a Q2 sería: D206 / D98 / D16, D100, D140, D442

EJERCICIO 3

Sea un Sistema de Recuperación de Información cuya colección consta de los siguientes documentos:

D1: El concepto de colección se ha transformado con el surgimiento de las bases de datos, que no necesitan de un espacio físico para ser accesibles a los usuarios.

D2: La biblioteconomía teórica incluye la teoría de la información y la gestión del conocimiento, así como los factores externos que influyen en ellas.

D3: La biblioteconomía aplicada se ocupa del desarrollo y mantenimiento de las colecciones, lo que implica diversos servicios técnicos, como adquisición, catalogación, préstamo o descarte.

D4: Puede entenderse el rol del bibliotecario como un intermediario entre el usuario y la colección de una unidad de información, asistiendo al usuario en el planteamiento y ejecución de su búsqueda de información.

D5: La organización de una biblioteca, esto es, la organización física de un catálogo de libros, ha llevado al estudio de cómo estructurar el conocimiento humano, área de estudio denominada clasificación.

El sistema ha procesado los textos de la colección eliminando las siguientes palabras, considerándolas palabras vacías: a, al, área, así, asistiendo, como, con, de, del, denominada, diversos, el, en, ellas, entenderse, entre, es, esto, estudio, estructurar, ha, implica, incluye, influyen, la, las, lo, los, llevado, necesitan, no, ocupa, para, puede, que, se, ser, su, surgimiento, transformado, un, una, y.

De igual forma, el sistema ha eliminado los acentos y las variantes de género y número de sustantivos y adjetivos, reuniendo todas las apariciones bajo las formas de masculino singular.

El sistema también ha efectuado la unificación de los siguientes términos relacionados semánticamente entre sí, mediante un algoritmo de stemming:

- Biblioteca, bibliotecario, biblioteconomía → bibliotec
- Catalogación, catálogo → catalog
- Teoría, teórico → teor

El SRI se basa en el modelo booleano extendido y el peso de los términos en los documentos se calcula como el cociente entre la frecuencia del término en el documento y la frecuencia máxima de un término en un documento cualquiera de la colección (en este caso, considere 2 como dicha frecuencia máxima).

Es suficiente utilizar dos decimales en los cálculos. A la hora de representar la respuesta, se indicará con el símbolo '/' el orden entre documentos, y una coma servirá para indicar que el sistema no puede imponer un orden entre documentos (debido a que poseen la misma similaridad con la consulta). Así, la respuesta D3 / D5 / D2, D4 / D1 indicaría que el sistema

situaría en primer lugar, con la máxima similaridad, el documento D3; a continuación, con menor similaridad, el documento D5; a continuación, con la misma similaridad –aunque menor que la que posee el documento D5-, el sistema incluye los documentos D2 y D4; por último, con la menor similaridad, el documento D1.

Hallar la respuesta del sistema en relación a los 1000 primeros documentos de la colección ante las consultas:

Q1 = (unidad AND información)

Q2 = (organización OR catalogación)

SOLUCIÓN

Podemos resumir la información sobre los términos y los documentos donde aparecen en la siguiente matriz de ocurrencias, incluyendo la información sobre su frecuencia absoluta de aparición (la ausencia de número indica un 0):

	D1	D2	D3	D4	D5
accesible	1				
adquisicion			1		
aplicado			1		
base	1				
bibliotec		1	1	1	1
busqueda				1	
catalog			1		1
clasificacion					1
coleccion	1		1	1	
concepto	1				
conocimiento		1			1
dato	1				
desarrollo			1		
descarte			1		
ejecucion				1	
espacio	1				
externo		1			
factor		1			
fisico	1				1
gestion		1			
humano					1
informacion		1		2	
intermediario				1	
libro					1
mantenimiento			1		
organizacion					2
planteamiento				1	
prestamo			1		
rol				1	
servicio			1		
tecnico			1		
teor		2			
usuario	1			2	
unidad				1	

Dividiendo cada una de las frecuencias por 2 obtenemos los pesos de cada término en los documentos de la colección. Como en las consultas del ejercicio solo se emplean los términos unidad, informacion, organizacion y catalog, en la tabla resultante solo incluimos la información relativa a dichos términos:

	N	LISTA DOCUMENTOS	LISTA APARICIONES (NÚMDOC, PESO, <POSIC>)
unidad	1	4	(4,0'5,<18>)
informacion	2	2, 4	(2,0'5,<9>), (4,1,<20,33>)
organizacion	1	5	(5,1,<2,9>)
catalog	2	3, 5	(3,0'5,<21>), (5,0'5,<13>),

Q1 = unidad AND informacion

Calculamos la similaridad correspondiente a la conectiva AND entre la consulta Q1 y cada uno de los 5 documentos de la colección, considerando en el eje X los pesos de los documentos en relación al término 'unidad' y en el eje Y los pesos de los documentos en relación al término 'informacion':

$SIM_{AND}(Q1, D1) = 1 - \sqrt{\frac{(1-D1_{unidad})^2+(1-D1_{informacion})^2}{2}} = 1 - \sqrt{\frac{(1-0)^2+(1-0)^2}{2}} = 0$

$SIM_{AND}(Q1, D2) = 1 - \sqrt{\frac{(1-D2_{unidad})^2+(1-D2_{informacion})^2}{2}} = 1 - \sqrt{\frac{(1-0)^2+(1-0{,}5)^2}{2}} = 0'21$

$SIM_{AND}(Q1, D3) = 1 - \sqrt{\frac{(1-D3_{unidad})^2+(1-D3_{informacion})^2}{2}} = 1 - \sqrt{\frac{(1-0)^2+(1-0)^2}{2}} = 0$

$SIM_{AND}(Q1, D4) = 1 - \sqrt{\frac{(1-D4_{unidad})^2+(1-D4_{informacion})^2}{2}} = 1 - \sqrt{\frac{(1-0{,}5)^2+(1-1)^2}{2}} = 0'64$

$SIM_{AND}(Q1, D5) = 1 - \sqrt{\frac{(1-D5_{unidad})^2+(1-D5_{informacion})^2}{2}} = 1 - \sqrt{\frac{(1-0)^2+(1-0)^2}{2}} = 0$

Por tanto, la respuesta del Sistema a Q1 sería la siguiente:

D4 / D2 / D1, D3, D5

Q2 = organización OR catalogación

Calculamos la similaridad correspondiente a la conectiva OR entre la consulta Q2 y cada uno de los 5 documentos de la colección, considerando en el eje X los pesos de los documentos en relación al término 'organizacion' y en el eje Y los pesos de los documentos en relación al término 'catalog':

$SIM_{OR}(Q2, D1) = \sqrt{\frac{D1^2_{organizacion} + D1^2_{catalog}}{2}} = \sqrt{\frac{0^2+0^2}{2}} = 0$

$SIM_{OR}(Q2, D2) = \sqrt{\frac{D2^2_{organizacion} + D2^2_{catalog}}{2}} = \sqrt{\frac{0^2+0^2}{2}} = 0$

$SIM_{OR}(Q2, D3) = \sqrt{\frac{D3^2_{organizacion} + D3^2_{catalog}}{2}} = \sqrt{\frac{0^2+0{,}5^2}{2}} = 0'36$

$SIM_{OR}(Q2, D4) = \sqrt{\frac{D4^2_{organizacion} + D4^2_{catalog}}{2}} = \sqrt{\frac{0^2+0^2}{2}} = 0$

$SIM_{OR}(Q2, D5) = \sqrt{\frac{D5^2_{organizacion} + D5^2_{catalog}}{2}} = \sqrt{\frac{1^2+0{,}5^2}{2}} = 0'79$

Por tanto, la respuesta del Sistema a Q2 sería la siguiente:

D5 / D3 / D1, D2, D4

En resumen, la respuesta a Q1 sería: D4 / D2 / D1, D3, D5

Y la respuesta del Sistema a Q2 sería: D5 / D3 / D1, D2, D4

EJERCICIO 4

Sea un Sistema de Recuperación de Información cuya matriz de ocurrencias término/documento es la siguiente, incluyendo la información sobre su frecuencia absoluta de aparición (solamente los seis primeros documentos y los doce primeros términos):

	D1	D2	D3	D4	D5	D6
base	1	0	0	2	0	1
banco	2	1	0	0	0	0
dato	1	1	1	0	0	1
sistema	0	0	1	1	0	0
gestor	0	1	2	0	0	1
SGBD	3	1	0	0	2	1
DBMS	2	1	0	2	2	3
estatica	0	2	1	0	0	0
documental	1	0	1	0	1	0
consulta	0	0	3	1	1	0
modelo	0	1	0	1	0	0
multidimensional	1	0	0	2	0	0

El SRI se basa en el modelo booleano extendido y el peso de los términos en los documentos se calcula como el cociente entre la frecuencia del término en el documento y la frecuencia máxima de un término en un documento cualquiera de la colección (en este caso, considere 3 como dicha frecuencia máxima).

Es suficiente utilizar dos decimales en los cálculos. A la hora de representar la respuesta, se indicará con el símbolo '/' el orden entre documentos, y una coma servirá para indicar que el sistema no puede imponer un orden entre documentos (debido a que poseen la misma similaridad con la consulta). Así, la respuesta D3 / D5 / D2, D4 / D1 indicaría que el sistema situaría en primer lugar, con la máxima similaridad, el documento D3; a continuación, con menor similaridad, el documento D5; a continuación, con la misma similaridad –aunque menor que la que posee el documento D5-, el sistema incluye los documentos D2 y D4; por último, con la menor similaridad, el documento D1.

Hallar la respuesta del sistema en relación a los seis primeros documentos de la colección ante las consultas:

Q1 = (base AND dato)

Q2 = (consulta OR DBMS)

SOLUCIÓN

Dividiendo cada una de las frecuencias por 3 obtenemos los pesos de cada término en los documentos de la colección. Como en las consultas del ejercicio solo se emplean los términos base, dato, consulta y DBMS, en la tabla resultante solo incluimos la información relativa a dichos términos:

	N	LISTA DOCUMENTOS	LISTA APARICIONES (NÚMDOC, PESO, <POSIC>)
base	3	1, 4, 6	(1,0'33,<18>), (4,0'67,<2,36>), (6,0'33,<57>)
dato	4	1, 2, 3, 6	(1,0'33,<9>), (2,0'33,<25>), (3,0'33,<40>), (6,0'33,<12>)
consulta	3	3, 4, 5	(3,1,<2,9,11>), (4,0'33,<98>), (5,0'33,<41>)
DBMS	5	1, 2, 4, 5, 6	(1,0'67,<21,76>), (2,0'33,<13>), (4,0'67,<3,8>), (5,0'67,<73,91>), (6,1,<2,9,14>)

Q1 = base AND dato

Calculamos la similaridad correspondiente a la conectiva AND entre la consulta Q1 y cada uno de los 6 documentos de la colección, considerando en el eje X los pesos de los documentos en relación al término 'base' y en el eje Y los pesos de los documentos en relación al término 'dato':

$SIM_{AND}(Q1, D1) = 1 - \sqrt{\frac{(1-D1_{base})^2+(1-D1_{dato})^2}{2}} = 1 - \sqrt{\frac{(1-0,33)^2+(1-0,33)^2}{2}} = 0'33$

$SIM_{AND}(Q1, D2) = 1 - \sqrt{\frac{(1-D2_{base})^2+(1-D2_{dato})^2}{2}} = 1 - \sqrt{\frac{(1-0)^2+(1-0,33)^2}{2}} = 0'15$

$SIM_{AND}(Q1, D3) = 1 - \sqrt{\frac{(1-D3_{base})^2+(1-D3_{dato})^2}{2}} = 1 - \sqrt{\frac{(1-0)^2+(1-0,33)^2}{2}} = 0'15$

$SIM_{AND}(Q1, D4) = 1 - \sqrt{\frac{(1-D4_{base})^2+(1-D4_{dato})^2}{2}} = 1 - \sqrt{\frac{(1-0,67)^2+(1-0)^2}{2}} = 0'25$

$SIM_{AND}(Q1, D5) = 1 - \sqrt{\frac{(1-D5_{base})^2+(1-D5_{dato})^2}{2}} = 1 - \sqrt{\frac{(1-0)^2+(1-0)^2}{2}} = 0$

$SIM_{AND}(Q1, D6) = 1 - \sqrt{\frac{(1-D6_{base})^2+(1-D6_{dato})^2}{2}} = 1 - \sqrt{\frac{(1-0,33)^2+(1-0,33)^2}{2}} = 0'33$

Por tanto, la respuesta del Sistema a Q1 sería la siguiente:

D1, D6 / D4 / D2, D3 / D5

Q2 = consulta OR DBMS

Calculamos la similaridad correspondiente a la conectiva OR entre la consulta Q2 y cada uno de los 6 documentos de la colección, considerando en el eje X los pesos de los documentos en relación al término 'consulta' y en el eje Y los pesos de los documentos en relación al término 'DBMS':

$SIM_{OR}(Q2, D1) = \sqrt{\frac{D1^2_{consulta}+D1^2_{DBMS}}{2}} = \sqrt{\frac{0^2+0{,}67^2}{2}} = 0'48$

$SIM_{OR}(Q2, D2) = \sqrt{\frac{D2^2_{consulta}+D2^2_{DBMS}}{2}} = \sqrt{\frac{0^2+0{,}33^2}{2}} = 0'23$

$SIM_{OR}(Q2, D3) = \sqrt{\frac{D3^2_{consulta}+D3^2_{DBMS}}{2}} = \sqrt{\frac{1^2+0^2}{2}} = 0'71$

$SIM_{OR}(Q2, D4) = \sqrt{\frac{D4^2_{consulta}+D4^2_{DBMS}}{2}} = \sqrt{\frac{0{,}33^2+0{,}67^2}{2}} = 0'53$

$SIM_{OR}(Q2, D5) = \sqrt{\frac{D5^2_{consulta}+D5^2_{DBMS}}{2}} = \sqrt{\frac{0{,}33^2+0{,}67^2}{2}} = 0'53$

$SIM_{OR}(Q2, D6) = \sqrt{\frac{D6^2_{consulta}+D6^2_{DBMS}}{2}} = \sqrt{\frac{0^2+1^2}{2}} = 0'71$

Por tanto, la respuesta del Sistema a Q2 sería la siguiente:

D3, D6 / D4, D5 / D1 / D2

En resumen, la respuesta a Q1 sería: D1, D6 / D4 / D2, D3 / D5

Y la respuesta del Sistema a Q2 sería: D3, D6 / D4, D5 / D1 / D2

EJERCICIO 5

Sea un Sistema de Recuperación de Información cuya colección consta de un millón de documentos. La información relativa a los primeros cinco documentos de dicha colección (los términos de indexación que contiene, su frecuencia de aparición y sus posiciones dentro de los documentos) se resume en la siguiente tabla:

DOCUMENTOS	TÉRMINOS (FRECUENCIA, <POSICIONES>)
D1	planificacion(2,<15,62>), sistema(1,<20>), automatizacion(1,<22>), biblioteca(1,<4>), programa(1,<37>), informatico(1,<38>), proceso(1,<57>)
D2	demostracion(1,<49>), producto(2,<7,52>), practica(1,<50>), sistema(1,<11>), gestion(1,<13>), biblioteca(1,<14>), tratamiento(1,<6>), texto(1,<8>), imagen(1,<10>)
D3	tratamiento(1,<91>), automatizacion(1,<92>), fuentes(1,<153>), recursos(1,<155>), texto(1,<52>), foto(1,<53>), pagina(1,<55>), web(1,<56>), archivo(1,<70>)
D4	recurso(1,<26>), digital(1,<27>), aplicacion(1,<48>), informatico(1,<49>), automatizacion(1,<73>),archivo(1,<90>)
D5	centro(1,<35>), documentacion(1,<37>), archivo(2,<72,87>), digital(1,<88>), gestion(1,<105>), biblioteca(1,<106>)

El SRI se basa en el modelo booleano extendido y el peso de los términos en los documentos se calcula como el cociente entre la frecuencia del término en el documento y la frecuencia máxima de un término en un documento cualquiera de la colección (en este caso, considere 2 como dicha frecuencia máxima).

Es suficiente utilizar dos decimales en los cálculos. A la hora de representar la respuesta, se indicará con el símbolo '/' el orden entre documentos, y una coma servirá para indicar que el sistema no puede imponer un orden entre documentos (debido a que poseen la misma similaridad con la consulta). Así, la respuesta D3 / D5 / D2, D4 / D1 indicaría que el sistema situaría en primer lugar, con la máxima similaridad, el documento D3; a continuación, con menor similaridad, el documento D5; a continuación, con la misma similaridad –aunque menor que la que posee el documento D5-, el sistema incluye los documentos D2 y D4; por último, con la menor similaridad, el documento D1.

Hallar la respuesta del sistema en relación a los cinco primeros documentos de la colección ante las consultas:

Q1 = (archivo AND digital)

Q2 = (gestion OR planificacion)

SOLUCIÓN

Dividiendo cada una de las frecuencias por 2 obtenemos los pesos de cada término en los documentos de la colección. Como en las consultas del ejercicio solo se emplean los términos archivo, digital, gestion y planificacion, en la tabla resultante solo incluimos la información relativa a dichos términos:

	N	LISTA DOCUMENTOS	LISTA APARICIONES (NÚMDOC, PESO, <POSIC>)
archivo	3	3, 4, 5	(3,0'5,<70>), (4,0'5,<90>), (5,1,<72,87>)
digital	2	4, 5	(4,0'5,<27>), (5,0'5,<88>)
gestion	2	2, 5	(2,0'5,<13>), (5,0'5,<105>)
planificacion	1	1	(1,1,<15,62>)

Q1 = archivo AND digital

Calculamos la similaridad correspondiente a la conectiva AND entre la consulta Q1 y cada uno de los 5 documentos de la colección, considerando en el eje X los pesos de los documentos en relación al término 'archivo' y en el eje Y los pesos de los documentos en relación al término 'digital':

$SIM_{AND}(Q1, D1) = 1 - \sqrt{\frac{(1-D1_{archivo})^2+(1-D1_{digital})^2}{2}} = 1 - \sqrt{\frac{(1-0)^2+(1-0)^2}{2}} = 0$

$SIM_{AND}(Q1, D2) = 1 - \sqrt{\frac{(1-D2_{archivo})^2+(1-D2_{digital})^2}{2}} = 1 - \sqrt{\frac{(1-0)^2+(1-0)^2}{2}} = 0$

$SIM_{AND}(Q1, D3) = 1 - \sqrt{\frac{(1-D3_{archivo})^2+(1-D3_{digital})^2}{2}} = 1 - \sqrt{\frac{(1-0,5)^2+(1-0)^2}{2}} = 0'21$

$SIM_{AND}(Q1, D4) = 1 - \sqrt{\frac{(1-D4_{archivo})^2+(1-D4_{digital})^2}{2}} = 1 - \sqrt{\frac{(1-0,5)^2+(1-0,5)^2}{2}} = 0'5$

$SIM_{AND}(Q1, D5) = 1 - \sqrt{\frac{(1-D5_{archivo})^2+(1-D5_{digital})^2}{2}} = 1 - \sqrt{\frac{(1-1)^2+(1-0,5)^2}{2}} = 0'64$

Por tanto, la respuesta del Sistema a Q1 sería la siguiente:

D5 / D4 / D3 / D1, D2

Q2 = gestion OR planificación

Calculamos la similaridad correspondiente a la conectiva OR entre la consulta Q2 y cada uno de los 5 documentos de la colección, considerando en el eje X los pesos de los documentos en relación al término 'gestion' y en el eje Y los pesos de los documentos en relación al término 'planificacion':

$SIM_{OR}(Q2, D1) = \sqrt{\frac{D1^2_{gestion}+D1^2_{planificacion}}{2}} = \sqrt{\frac{0^2+1^2}{2}} = 0'71$

$SIM_{OR}(Q2, D2) = \sqrt{\frac{D2^2_{gestion}+D2^2_{planificacion}}{2}} = \sqrt{\frac{0{,}5^2+0^2}{2}} = 0'35$

$SIM_{OR}(Q2, D3) = \sqrt{\frac{D3^2_{gestion}+D3^2_{planificacion}}{2}} = \sqrt{\frac{0^2+0^2}{2}} = 0$

$SIM_{OR}(Q2, D4) = \sqrt{\frac{D4^2_{gestion}+D4^2_{planificacion}}{2}} = \sqrt{\frac{0^2+0^2}{2}} = 0$

$SIM_{OR}(Q2, D5) = \sqrt{\frac{D5^2_{gestion}+D5^2_{planificacion}}{2}} = \sqrt{\frac{0{,}5^2+0^2}{2}} = 0'35$

Por tanto, la respuesta del Sistema a Q2 sería la siguiente:

D1 / D2, D5 / D3, D4

En resumen, la respuesta a Q1 sería: D5 / D4 / D3 / D1, D2

Y la respuesta del Sistema a Q2 sería: D1 / D2, D5 / D3, D4

EJERCICIO 6

Sea un Sistema de Recuperación de Información cuya colección consta de un millón de documentos. La colección incluye 500000 términos de indexación, de los cuales los términos incluidos en la consulta tienen la siguiente distribución:

TÉRMINOS	N
t11	696174
t209	286
t34815	75903
t487161	142605

Siendo 'n' el número de documentos de la colección en los que aparece un determinado término.

Seis de los documentos de la colección están representados (en relación a los términos de la consulta) de la siguiente manera (incluyendo la frecuencia de aparición en cada documento):

DOCUMENTOS	T11	T209	T34815	T487161
D1	2	0	0	3
D428	3	1	0	1
D49067	0	1	0	2
D102314	1	0	0	1
D624752	0	2	0	1
D991023	0	1	3	0

El SRI se basa en el modelo booleano extendido y el peso de los términos en los documentos se calcula como el cociente entre la frecuencia del término en el documento y la frecuencia máxima de un término en un documento cualquiera de la colección (en este caso, considere 3 como dicha frecuencia máxima).

Es suficiente utilizar dos decimales en los cálculos. A la hora de representar la respuesta, se indicará con el símbolo '/' el orden entre documentos, y una coma servirá para indicar que el sistema no puede imponer un orden entre documentos (debido a que poseen la misma similaridad con la consulta). Así, la respuesta D3 / D5 / D2, D4 / D1 indicaría que el sistema situaría en primer lugar, con la máxima similaridad, el documento D3; a continuación, con menor similaridad, el documento D5; a continuación, con la misma similaridad –aunque menor que la que posee el documento D5-, el sistema incluye los documentos D2 y D4; por último, con la menor similaridad, el documento D1.

Hallar la respuesta del sistema en relación a los seis documentos de la colección indicados anteriormente y ante las consultas:

Q1 = (t209 AND t487161)

Q2 = (t34815 OR t11)

SOLUCIÓN

Dividiendo cada una de las frecuencias por 3 obtenemos los pesos de cada término en los documentos de la colección. Como en las consultas del ejercicio se emplean los cuatro términos, en la tabla resultante incluimos la información relativa a todos los términos:

DOCUMENTOS	T11	T209	T34815	T487161
D1	0'67	0	0	1
D428	1	0'33	0	0'33
D49067	0	0'33	0	0'67
D102314	0'33	0	0	0'33
D624752	0	0'67	0	0'33
D991023	0	0'33	1	0

Q1 = t209 AND t487161

Calculamos la similaridad correspondiente a la conectiva AND entre la consulta Q1 y cada uno de los 6 documentos de la colección indicados, considerando en el eje X los pesos de los documentos en relación al término 't209' y en el eje Y los pesos de los documentos en relación al término 't487161':

$SIM_{AND}(Q1, D1) = 1 - \sqrt{\frac{(1-D1_{t209})^2+(1-D1_{t487161})^2}{2}} = 1 - \sqrt{\frac{(1-0)^2+(1-1)^2}{2}} = 0'29$

$SIM_{AND}(Q1, D428) = 1 - \sqrt{\frac{(1-D428_{t209})^2+(1-D428_{t487161})^2}{2}} = 1 - \sqrt{\frac{(1-0,33)^2+(1-0,33)^2}{2}} = 0'33$

$SIM_{AND}(Q1, D49067) = 1 - \sqrt{\frac{(1-D49067_{t209})^2+(1-D49067_{t487161})^2}{2}} = 1 - \sqrt{\frac{(1-0,33)^2+(1-0,67)^2}{2}} =$ 0'47

$SIM_{AND}(Q1, D102314) = 1 - \sqrt{\frac{(1-D102314_{t209})^2+(1-D102314_{t487161})^2}{2}} = 1 - \sqrt{\frac{(1-0)^2+(1-0,33)^2}{2}} =$ 0'15

$SIM_{AND}(Q1, D624752) = 1 - \sqrt{\frac{(1-D624752_{t209})^2+(1-D624752_{t487161})^2}{2}} = 1 - \sqrt{\frac{(1-0,67)^2+(1-0,33)^2}{2}} =$ 0'47

$SIM_{AND}(Q1, D991023) = 1 - \sqrt{\frac{(1-D991023_{t209})^2+(1-D991023_{t487161})^2}{2}} = 1 - \sqrt{\frac{(1-0,33)^2+(1-0)^2}{2}} =$ 0'15

Por tanto, la respuesta del Sistema a Q1 sería la siguiente:

D49067, D624752 / D428 / D1 / D102314, D991023

Q2 = t34815 OR t11

Calculamos la similaridad correspondiente a la conectiva OR entre la consulta Q2 y cada uno de los 6 documentos de la colección indicados, considerando en el eje X los pesos de los documentos en relación al término 't34815' y en el eje Y los pesos de los documentos en relación al término 't11':

$SIM_{OR}(Q2, D1) = \sqrt{\frac{D1_{t34815}^2 + D1_{t11}^2}{2}} = \sqrt{\frac{0^2 + 0{,}67^2}{2}} = 0'48$

$SIM_{OR}(Q2, D428) = \sqrt{\frac{D428_{t34815}^2 + D428_{t11}^2}{2}} = \sqrt{\frac{0^2 + 1^2}{2}} = 0'71$

$SIM_{OR}(Q2, D49067) = \sqrt{\frac{D49067_{t34815}^2 + D49067_{t11}^2}{2}} = \sqrt{\frac{0^2 + 0^2}{2}} = 0$

$SIM_{OR}(Q2, D102314) = \sqrt{\frac{D102314_{t34815}^2 + D102314_{t11}^2}{2}} = \sqrt{\frac{0^2 + 0{,}33^2}{2}} = 0'23$

$SIM_{OR}(Q2, D624752) = \sqrt{\frac{D624752_{t34815}^2 + D624752_{t11}^2}{2}} = \sqrt{\frac{0^2 + 0^2}{2}} = 0$

$SIM_{OR}(Q2, D991023) = \sqrt{\frac{D991023_{t34815}^2 + D991023_{t11}^2}{2}} = \sqrt{\frac{1^2 + 0^2}{2}} = 0'71$

Por tanto, la respuesta del Sistema a Q2 sería la siguiente:

D428, D991023 / D1 / D102314 / D49067, D624752

En resumen, la respuesta a Q1 sería: D49067, D624752 / D428 / D1 / D102314, D991023

Y la respuesta del Sistema a Q2 sería: D428, D991023 / D1 / D102314 / D49067, D624752

MODELO BM25/Okapi

El modelo BM25 sigue los principios esenciales del modelo probabilístico clásico, llamado Binary Independence Model o BIM, aunque trata de superar sus limitaciones, en especial las relativas al peso de los términos y la normalización de la longitud de los documentos.

El modelo probabilístico clásico (Binary Independence Model o BIM) es capaz de ordenar los documentos de la respuesta por orden de su probabilidad de relevancia en relación a la consulta, pero se basa para ello en un esquema binario en cuanto al peso de los términos en los documentos y en las consultas. Ello implica considerar exclusivamente la presencia (mediante el número 1) o la ausencia (mediante el número 0) de los términos, sin poder reflejar la mayor o menor importancia de cada término en la descripción del contenido de los documentos. Una forma habitual de reflejar dicha importancia relativa consiste en emplear la frecuencia de aparición del término en los documentos, considerando que cuantas más veces aparezca un cierto término en un documento, en mayor medida el contenido de dicho documento puede ser descrito mediante ese término. Así pues, si la frecuencia de aparición de los términos aporta una información valiosa para calcular su relevancia en relación a una consulta, el esquema binario en el que se basa el modelo probabilístico clásico posee una limitación que conviene superar.

De igual modo, el esquema binario del modelo probabilístico clásico fuerza la consideración únicamente de la presencia (mediante el 1) o la ausencia (mediante el 0) de los términos en la consulta. La ponderación de los términos en la consulta mediante números reales permitiría al usuario cuantificar la importancia relativa de los términos de la consulta, reflejando con mayor precisión la necesidad informativa del usuario. En consecuencia, la ponderación binaria de los términos en las consultas supone una faceta más de la limitación del modelo clásico que convendría dominar.

El Modelo de Independencia Binaria clásico considera también que todos los documentos de la colección son de la misma longitud, de manera que no es necesario corregir las posibles diferencias que existan en el sistema en este aspecto. Sin embargo, cuanto más largo es un documento más palabras contiene, con lo que la probabilidad de que un documento extenso incluya alguno de los términos presentes en la consulta es mayor que si el documento es breve aun cuando aborden idéntica temática, siendo también mayor la probabilidad de que cada término esté presente más veces en un documento largo que en un documento corto. Se deduce, pues, la conveniencia de normalizar la longitud de los documentos, de manera que la longitud no sea un factor que favorezca la aparición de los documentos extensos en mejor posición que los documentos breves, cuando su relevancia temática en relación a una consulta es semejante.

Robertson y su equipo se proponen desarrollar un nuevo modelo probabilístico que supere las tres limitaciones expuestas: el esquema binario en los términos de los documentos, el esquema binario en los términos de las consultas y la falta de normalización de los documentos de la colección. Surgirán así los distintos modelos Best Match (BM) frente al modelo BIM inicial y el sistema experimental Okapi basado en dichos nuevos modelos.

Los modelos BM/Okapi incorporan la información relativa a la frecuencia de los términos en los documentos y en las consultas desde un punto de vista probabilístico mediante la distribución 2-Poisson. Conforme a esta distribución de Poisson, el número de ocurrencias de un término en los documentos, considerando sus valores medios, difiere según que el documento verse o no sobre el concepto representado por los términos de la consulta. De esta manera, la frecuencia de aparición de un término en un documento puede ser un indicador de su contenido:

- Si la frecuencia es cercana a la media Poisson de los documentos denominados élite, es decir, aquellos que versan sobre ese concepto, es más probable que el contenido del documento verse sobre dicho concepto.
- Al contrario, si la frecuencia es cercana a la media Poisson de los documentos no-élite, es decir, los que no versan sobre ese concepto, es más probable que el contenido del documento no verse sobre dicho concepto.

La dificultad de los modelos BM surge, al igual que en el modelo clásico, en la aplicación de la fórmula teórica a la hora de pesar los términos, pues se carece de conocimiento previo sobre las medias de ocurrencia en documentos élite y no-élite, así como tampoco conocemos los conceptos tratados por cada uno de los documentos.

De nuevo se recurre a la experimentación para evaluar distintas aproximaciones de las fórmulas de pesado. Estas distintas simplificaciones dan lugar a distintos modelos BM (BM1, BM11…), hasta llegar finalmente al modelo BM25, la versión más efectiva conseguida, cuyas fórmulas estándares de pesado se detallan a continuación.

El pesado correspondiente a los términos en las consultas sigue la fórmula siguiente:

$$w_{tq} = \frac{(k_3 + 1)\, tf_{tq}}{k_3 + tf_{tq}}$$

Donde:

- w_{tq} es el peso correspondiente al término ‘t’ en la consulta ‘q’
- tf_{tq} es la frecuencia de aparición del término ‘t’ en la consulta ‘q’
- k_3 es un parámetro mayo o igual que cero ($k_3 \geq 0$), cuyo valor depende de la colección y debe ser determinada experimentalmente

La fórmula del pesado de los términos en los documentos de la colección incluye también un factor ‘b’ para contrarrestar el efecto de la longitud de los documentos:

$$w_{td} = \frac{(k_1 + 1) tf_{td}}{k_1((1 - b) + b.\frac{dl_d}{avgdl}) + tf_{td}}$$

Donde:

- w_{td} es el peso correspondiente al término ‘t’ en el documento ‘d’
- tf_{td} es la frecuencia de aparición del término ‘t’ en el documento ‘d’

- k_1 es un parámetro mayor o igual que cero ($k_1 \geq 0$), cuyo valor depende de la colección y debe ser determinada experimentalmente
- El parámetro b ($0 \leq b \leq 1$) depende de la colección y debe ser determinado experimentalmente
- dl_d es la longitud en bytes del documento 'd'
- avgdl es la longitud media en bytes de los documentos de la colección

La fórmula final de la similaridad entre un documento 'd' y una consulta 'q' incluye, como en el modelo clásico, la posibilidad de incluir retroalimentación por relevancia:

$$Sim(d,q) = \sum_{t \in d\ y\ q} log \frac{(r + 0'5)/(R - r + 0'5)}{(n - r + 0'5)/(N - n - R + r + 0'5)} \cdot w_{td} \cdot w_{tq}$$

Donde:

- Sim(d,q) es el valor de similaridad entre el documento 'd' y la consulta 'q'
- w_{td} es el peso correspondiente al término 't' en el documento 'd'
- w_{tq} es el peso correspondiente al término 't' en la consulta 'q'
- R es el número de documentos relevantes en relación a la consulta 'q'
- r es el número de documentos relevantes que contienen el término 't'
- N es el número de documentos de la colección
- n es el número de documentos de la colección que contienen el término 't'

Si no se dispone de información facilitada por el usuario, por ejemplo al comienzo, entonces R=r=0 y esta componente se reduce a una variante del coeficiente de los términos empleada en la hipótesis inicial del modelo probabilístico clásico:

$$c(t) = log \frac{\frac{r + 0'5}{R - r + 0'5}}{\frac{n - r + 0'5}{N - n - R + r + 0'5}} = log \frac{1}{\frac{n + 0'5}{N - n + 0'5}}$$

$$c(t) = log \frac{N - n + 0'5}{n + 0'5}$$

Donde:

- N es el número de documentos de la colección
- N es el número de documentos de la colección que contienen el término 't'

Con el fin de comparar esta fórmula del coeficiente en el modelo BM25/Okapi con el valor del coeficiente en la hipótesis inicial del modelo probabilístico clásico, calcularemos dicho valor en el modelo BIM clásico. Para ello, recordemos que se asumían los siguientes valores:

$$p(t/R) = 0'5 \qquad p(t/\overline{R}) = \frac{n}{N}$$

Donde:

- $p(t/R)$ es la probabilidad de ocurrencia del término 't' en el conjunto de los documentos relevantes en relación a la consulta 'q'
- $p(t/\overline{R})$ es la probabilidad de ocurrencia del término 't' en el conjunto de los documentos irrelevantes en relación a la consulta 'q'
- N es el número de documentos de la colección
- n es el número de documentos de la colección que contienen el término 't'

$$c(t) = log\frac{p(t/R)}{1-p(t/R)} + log\frac{1-p(t/\overline{R})}{p(t/\overline{R})} = log\frac{0'5}{(1-0'5)} + log\frac{1-\frac{n}{N}}{\frac{n}{N}} = log\frac{\frac{N-n}{N}}{\frac{n}{N}}$$

$$c(t) = \log\frac{N-n}{n}$$

Como podemos comprobar, el valor del coeficiente es prácticamente el mismo en ambos casos, con la salvedad de que se suma 0'5 tanto al numerador como al denominador en el modelo BM25. Conviene destacar la semejanza de estas fórmulas con una de las variantes de la componente IDF del modelo vectorial:

$$IDF(término\ t) = log\frac{(N+1)}{n}$$

No se trata de parecidos casuales. Más adelante tendremos ocasión de observar la similitud de este modelo BM25 con los modelos vectoriales que efectúan normalización en la longitud de los documentos.

La fórmula de la similaridad entre un documento 'd' y una consulta 'q' cuando no se tienen datos procedentes de la retroalimentación (al comienzo del proceso, por ejemplo) es la siguiente:

$$Sim(d,q) = \sum_{t\in d\ y\ q} log\frac{N-n+0'5}{n+0'5} \,.\, w_{td} \,.\, w_{tq}$$

Donde:

- Sim(d,q) es el valor de similaridad entre el documento ‘d’ y la consulta ‘q’ cuando no se tienen datos procedentes de la retroalimentación (al comienzo del proceso, por ej.)
- w_{td} es el peso correspondiente al término ‘t’ en el documento ‘d’
- w_{tq} es el peso correspondiente al término ‘t’ en la consulta ‘q’
- N es el número de documentos de la colección
- n es el número de documentos de la colección que contienen el término ‘t’

Como hemos expuesto anteriormente, las fórmulas de los pesos empleadas en este modelo son las siguientes:

$$w_{td} = \frac{(k_1 + 1)tf_{td}}{k_1((1 - b) + b.\frac{dl_d}{avgdl}) + tf_{td}}$$

$$w_{tq} = \frac{(k_3 + 1)\, tf_{tq}}{k_3 + tf_{tq}}$$

En conclusión, la fórmula final de la similaridad entre un documento ‘d’ y una consulta ‘q’ cuando no se tienen datos procedentes de la retroalimentación (al comienzo del proceso, por ejemplo) es la siguiente:

$$Sim(d,q) = \sum_{t \in d\ y\ q} log\frac{N - n + 0'5}{n + 0'5} \cdot \frac{(k_1 + 1)tf_{td}}{k_1((1 - b) + b.\frac{dl_d}{avgdl}) + tf_{td}} \cdot \frac{(k_3 + 1)\, tf_{tq}}{k_3 + tf_{tq}}$$

Donde:

- Sim(d,q) es el valor de similaridad entre el documento ‘d’ y la consulta ‘q’ cuando no se tienen datos procedentes de la retroalimentación (al comienzo del proceso, por ej.)
- N es el número de documentos de la colección
- n es el número de documentos de la colección que contienen el término ‘t’
- k_1 es un parámetro mayor o igual que cero ($k_1 \geq 0$), cuyo valor depende de la colección y debe ser determinada experimentalmente
- tf_{td} es la frecuencia de aparición del término ‘t’ en el documento ‘d’
- El parámetro b ($0 \leq b \leq 1$) depende de la colección y debe ser determinado experimentalmente
- dl_d es la longitud en bytes del documento ‘d’
- avgdl es la longitud media en bytes de los documentos de la colección
- k_3 es un parámetro mayo o igual que cero ($k_3 \geq 0$), cuyo valor depende de la colección y debe ser determinada experimentalmente
- tf_{tq} es la frecuencia de aparición del término ‘t’ en la consulta ‘q’

Esta fórmula de la similaridad muestra la semejanza del modelo probabilístico BM25 con los modelos vectoriales que efectúan normalización de la longitud de los documentos. Puede comprobarse dicha similitud consultando, por ejemplo, la fórmula de la similaridad en el modelo vectorial con normalización por longitud basada en pivote (a continuación, en este mismo documento).

Los valores más usuales de los parámetros k_1, k_3 y b, cuya eficacia ha sido probada experimentalmente, son los siguientes:

- $k_1 = 1'2$
- $k_3 = 0$
- $b = 0'75$

En general, los valores de k_1 y k_3 pequeños reducen el impacto de la frecuencia del término en el documento y en la consulta respectivamente.

De hecho, en relación al parámetro k_3, habitualmente tratamos con consultas cortas en las que la frecuencia de los términos es 1, por lo que podemos no tener en consideración los efectos de la frecuencia en las consultas, e imponer consiguientemente un valor de $k_3 = 0$. Con ello estamos indicando que no hay necesidad de modificar la similaridad con este factor, pues con este valor de $k_3 = 0$, siendo la frecuencia de los términos en las consultas la unidad, el valor de $w_{tq}=1$.

BIBLIOGRAFÍA

Losada, David E. Modelos de recuperación de información II. En recuperación de información: Un enfoque práctico y multidisciplinar. Madrid: RA-MA, 2011, pp. 295-358.

Robertson, S.E.; Walker, S. (1994). Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. Proceedings of the 17th ACM SIGIR Conference, SIGIR 1994, pp. 232-241.

Robertson, S.E.; Walker, S.; Jones, S.; Hancock-Beaulieu, M.M.; Gatford, M. (1994). Okapi at TREC-3. Proceedings of the 3rd Text Retrieval Conference, TREC 1994, pp. 109-126.

Robertson, S.E.; Zaragoza, H. The probabilistic relevance framework: BM25 and beyond. Foundations and Trends in Information Retrieval, 2009, 3(4): 333-389.

EJERCICIO 1

Sea un Sistema de Recuperación de Información cuya colección consta de dos millones de documentos. Las consultas realizadas por el último usuario incluían los términos 'implantacion', 'sistema', 'gestion', 'automatizada' y 'biblioteca'. Los datos estadísticos esenciales en relación a dichos términos se resumen en la siguiente tabla:

	N	LISTA
implantacion	13048	(11,2), (90,1)
sistema	49173	(54,1), (63,2), (71,1)
gestion	66948	(27,2), (38,1), (84,3), (99,1)
automatizada	82163	(19,1), (27,1), (84,2)
biblioteca	135842	(19,1), (54,2), (84,1), (90,1), (99,2)

Donde 'n' es el número de documentos de la colección en los que aparece el término. La lista incluye los números de los 100 primeros documentos de la colección que contienen cada término, así como la frecuencia de aparición del término en cada documento. Así, (79,3) indica que el término aparece 3 veces en el documento 79.

Considere los valores usuales para los parámetros k1, k3 y b:

k1 = 1'2; k3 = 0; b = 0'75

La longitud de los 100 primeros documentos de la colección en los que aparecen los términos de la consulta es la siguiente:

D19 = 36 bytes; D27 = 28 bytes; D38 = 40 bytes; D54 = 50 bytes; D84 = 25 bytes;

D90 = 30 bytes; D99 = 32 bytes

Considere que la longitud media de los documentos de la colección es de 30 bytes:

avgdl = 30 bytes

Tratándose del comienzo del proceso, considere que R = r = 0

Es suficiente utilizar dos decimales en los cálculos. A la hora de representar la respuesta, se indicará con el símbolo '/' el orden entre documentos, y una coma servirá para indicar que el sistema no puede imponer un orden entre documentos (debido a que poseen la misma similaridad con la consulta). Así, la respuesta D3 / D5 / D2, D4 / D1 indicaría que el sistema situaría en primer lugar, con la máxima similaridad, el documento D3; a continuación, con menor similaridad, el documento D5; a continuación, con la misma similaridad –aunque menor que la que posee el documento D5-, el sistema incluye los documentos D2 y D4; por último, con la menor similaridad, el documento D1.

Hallar la respuesta inicial del sistema en relación a los 100 primeros documentos de la colección ante la consulta:

Q = gestion automatizada biblioteca

SOLUCIÓN

Hallamos los valores de la constante K para los 100 primeros documentos de la colección donde aparecen los términos de la consulta:

K(D19) = k1((1-b) + b.dl_{D19}/avgdl) = 1’2 ((1-0’75) + 0’75.36/30) = 1’38

K(D27) = k1((1-b) + b.dl_{D27}/avgdl) = 1’2 ((1-0’75) + 0’75.28/30) = 1’14

K(D38) = k1((1-b) + b.dl_{D38}/avgdl) = 1’2 ((1-0’75) + 0’75.40/30) = 1’5

K(D54) = k1((1-b) + b.dl_{D54}/avgdl) = 1’2 ((1-0’75) + 0’75.50/30) = 1’8

K(D84) = k1((1-b) + b.dl_{D84}/avgdl) = 1’2 ((1-0’75) + 0’75.25/30) = 1’05

K(D90) = k1((1-b) + b.dl_{D90}/avgdl) = 1’2 ((1-0’75) + 0’75.30/30) = 1’2

K(D99) = k1((1-b) + b.dl_{D99}/avgdl) = 1’2 ((1-0’75) + 0’75.32/30) = 1’26

Resumimos los pesos de los términos en estos documentos en la siguiente tabla:

W_{td}

	D19	D27	D38	D54	D84	D90	D99
gestion	0	$\frac{2'2\,.\,2}{1'14+2}$	$\frac{2'2.1}{1'5+1}$	0	$\frac{2'2.3}{1'05+3}$	0	$\frac{2'2.1}{1'26+1}$
automatizada	$\frac{2'2.1}{1'38+1}$	$\frac{2'2.1}{1'14+1}$	0	0	$\frac{2'2.2}{1'05+2}$	0	0
biblioteca	$\frac{2'2.1}{1'38+1}$	0	0	$\frac{2'2.2}{1'8+2}$	$\frac{2'2.1}{1'05+1}$	$\frac{2'2.1}{1'2+1}$	$\frac{2'2.2}{1'26+2}$

Una vez calculados los valores correspondientes a cada casilla:

W_{td}

	D19	D27	D38	D54	D84	D90	D99
gestion	0	1’40	0’88	0	1’63	0	0’97
automatizada	0’92	1’03	0	0	1’44	0	0
biblioteca	0’92	0	0	1’16	1’07	1	1’35

Resumimos los pesos de los términos en la consulta en la siguiente tabla:

W_{tq}

GESTION	1
automatizada	1
biblioteca	1

Calculamos a continuación los valores de los coeficientes correspondientes a los términos de la consulta:

c(gestion) = $\log\frac{2000000-66948+0,5}{66948+0,5}$ = log 28’87 = 1’46

c(automatizada) = $\log\frac{2000000-82163+0,5}{82163+0,5}$ = log 23’34 = 1’37

c(biblioteca) = $\log\frac{2000000-135842+0,5}{135842+0,5}$ = log 13’72 = 1’14

Calculamos finalmente los valores de similaridad entre documentos y consulta:

sim(D19,Q) = c(automatizada).w(automatizada,D19).1 + c(biblioteca).w(biblioteca,D19) = 2’31

sim(D27,Q) = c(gestion).w(gestion,D27) + c(automatizada).w(automatizada,D27) = 3’52

sim(D38,Q) = c(gestion).w(gestion,D38).1 = 1’46 x 0’88 = 1’28

sim(D54,Q) = c(biblioteca).w(biblioteca,54).1 = 1’14 x 1’16 = 1’32

sim(D84,Q) = c(gestion).w(gestion,D84)+c(automat).w(automat,D84)+c(bib).w(bib,D84) = 7’00

sim(D90,Q) = c(biblioteca).w(biblioteca,D90).1 = 1’14 x 1 = 1’14

sim(D99,Q) = c(gestion).w(gestion,D99).1 + c(biblioteca).w(biblioteca,D99).1 = 3’45

El resto de los 100 primeros documentos de la colección tienen un valor de similaridad 0 con la consulta Q, por lo que los eliminamos de la respuesta.

Por tanto, la respuesta del Sistema a Q sería: D84 / D27 / D99 / D19 / D54 / D38 / D90

EJERCICIO 2

Sea un Sistema de Recuperación de Información cuya colección consta de 3 millones de documentos. Los datos estadísticos esenciales en relación a los principales términos de indexación empleados en las últimas consultas se resumen en la siguiente tabla:

	N	LISTA DE DOCS (ENTRE LOS 100 PRIMEROS)	LISTA DE FRECUENCIA Y POSICIONES
t1	65423	43,68	(43,1,<132>),(68,1,<286>)
t2	517399	3,11, 57, 84	(3,2,<44,209>),(11,1,<315>),(57,1,<102>),(84,2,<12,345>)
t3	1471863	2,11,62,77,90	(2,1,<169>),(11,1,<66>),(62,2,<43,137>),(77,1,<216>), (90,3,<74,132,209>)
t4	806137	24,36,62,77,93	(24,1,<9>),(36,1,<101>),(62,1,<224>),(77,2,<12,44>), (93,2,<331,542>)
t5	242649	11,68,70	(11,2,<5,403>),(68,1,<45>),(70,3,<21,46,67>)

Donde 'n' es el número de documentos de la colección en los que aparece el término. La lista de documentos incluye los números de los 100 primeros documentos de la colección que contienen cada término. La lista de frecuencia y posiciones incluye el número de documento donde surge el término, la frecuencia de aparición en dicho documento y sus posiciones absolutas. Así, (79,1,<40>) indica que el término aparece en el documento número 79, una sola vez, en la posición 40 del documento.

Considere los valores usuales para los parámetros k1, k3 y b:

$$k1 = 1'2;\ \ k3 = 0;\ \ b = 0'75$$

La longitud de los 100 primeros documentos de la colección en los que aparecen los términos de la consulta es la siguiente:

D2 = 36 bytes; D3 = 28 bytes; D11 = 40 bytes; D24 = 50 bytes; D36 = 25 bytes;

D57 = 30 bytes; D62 = 32 bytes; D77 = 20 bytes; D84 = 42bytes; D90 = 34 bytes; D93 = 38 bytes

Considere que la longitud media de los documentos de la colección es de 30 bytes:

$$avgdl = 30 \text{ bytes}$$

Tratándose del comienzo del proceso, considere que R = r = 0

Es suficiente utilizar dos decimales en los cálculos. A la hora de representar la respuesta, se indicará con el símbolo '/' el orden entre documentos, y una coma servirá para indicar que el sistema no puede imponer un orden entre documentos (debido a que poseen la misma similaridad con la consulta). Así, la respuesta D3 / D5 / D2, D4 / D1 indicaría que el sistema situaría en primer lugar, con la máxima similaridad, el documento D3; a continuación, con menor similaridad, el documento D5; a continuación, con la misma similaridad –aunque menor

que la que posee el documento D5-, el sistema incluye los documentos D2 y D4; por último, con la menor similaridad, el documento D1.

Hallar la respuesta inicial del sistema en relación a los 100 primeros documentos de la colección ante la consulta:

$$Q = t2\ \ t3\ \ t4$$

SOLUCIÓN

Hallamos los valores de la constante K para los 100 primeros documentos de la colección donde aparecen los términos de la consulta:

K(D2) = k1((1-b) + b.dl_{D2}/avgdl) = 1’2 ((1-0’75) + 0’75.36/30) = 1’38

K(D3) = k1((1-b) + b.dl_{D3}/avgdl) = 1’2 ((1-0’75) + 0’75.28/30) = 1’14

K(D11) = k1((1-b) + b.dl_{D11}/avgdl) = 1’2 ((1-0’75) + 0’75.40/30) = 1’5

K(D24) = k1((1-b) + b.dl_{D24}/avgdl) = 1’2 ((1-0’75) + 0’75.50/30) = 1’8

K(D36) = k1((1-b) + b.dl_{D36}/avgdl) = 1’2 ((1-0’75) + 0’75.25/30) = 1’05

K(D57) = k1((1-b) + b.dl_{D57}/avgdl) = 1’2 ((1-0’75) + 0’75.30/30) = 1’2

K(D62) = k1((1-b) + b.dl_{D62}/avgdl) = 1’2 ((1-0’75) + 0’75.32/30) = 1’26

K(D77) = k1((1-b) + b.dl_{D77}/avgdl) = 1’2 ((1-0’75) + 0’75.20/30) = 0’90

K(D84) = k1((1-b) + b.dl_{D84}/avgdl) = 1’2 ((1-0’75) + 0’75.42/30) = 1’56

K(D90) = k1((1-b) + b.dl_{D90}/avgdl) = 1’2 ((1-0’75) + 0’75.34/30) = 1’32

K(D93) = k1((1-b) + b.dl_{D93}/avgdl) = 1’2 ((1-0’75) + 0’75.38/30) = 1’44

Resumimos los pesos de los términos en estos documentos en la siguiente tabla:

W_{td}

	D2	D3	D11	D24	D36	D57	D62	D77	D84	D90	D93
t2	0	$\frac{2'2.2}{1'14+2}$	$\frac{2'2.1}{1'5+1}$	0	0	$\frac{2'2.1}{1'2+1}$	0	0	$\frac{2'2.2}{1'56+2}$	0	0
t3	$\frac{2'2.1}{1'38+1}$	0	$\frac{2'2.1}{1'5+1}$	0	0	0	$\frac{2'2.2}{1'26+2}$	$\frac{2'2.1}{0'9+1}$	0	$\frac{2'2.3}{1'32+3}$	0
t4	0	0	0	$\frac{2'2.1}{1'8+1}$	$\frac{2'2.1}{1'05+1}$	0	$\frac{2'2.1}{1'26+1}$	$\frac{2'2.2}{0'9+2}$	0	0	$\frac{2'2.2}{1'44+2}$

Una vez calculados los valores correspondientes a cada casilla:

W_{td}

	D2	D3	D11	D24	D36	D57	D62	D77	D84	D90	D93
t2	0	1’40	0’88	0	0	1	0	0	1’24	0	0
t3	0’92	0	0’85	0	0	0	1’35	1’16	0	1’53	0
t4	0	0	0	0’79	1’07	0	0’97	1’52	0	0	1’28

Resumimos los pesos de los términos en la consulta en la siguiente tabla:

W_{tq}

t2	1
t3	1
t4	1

Calculamos a continuación los valores de los coeficientes correspondientes a los términos de la consulta:

c(t2) = $\log\frac{3000000-517399+0,5}{517399+0,5}$ = log 4’80 = 0’68

c(t3) = $\log\frac{3000000-1471863+0,5}{1471863+0,5}$ = log 1’04 = 0’02

c(t4) = $\log\frac{3000000-806137+0,5}{806137+0,5}$ = log 2’72 = 0’43

Calculamos finalmente los valores de similaridad entre documentos y consulta:

sim(D2,Q) = c(t3).w(t3,D2).1 = 0’02 x 0’92 = 0’02

sim(D3,Q) = c(t2).w(t2,D3) = 0’68 x 1’40 = 0’95

sim(D11,Q) = c(t2).w(t2,D11).1+ c(t3).w(t3,D11).1 = 0’68 x 0’88 + 0’02 x 0’85 = 0’6 + 0’02 = 0’62

sim(D24,Q) = c(t4).w(t4,D24).1 = 0’43 x 0’79 = 0’34

sim(D36,Q) = c(t4).w(t4,D36) = 0’43 x 1’07 = 0’46

sim(D57,Q) = c(t2).w(t2,D57).1 = 0’68 x 1 = 0’68

sim(D62,Q) = c(t3).w(t3,D62).1 + c(t4).w(t4,D62).1 = 0’02 x 1’35 + 0’43 x 0’97=0’03 + 0’42= 0’45

sim(D77,Q) = c(t3).w(t3,D77).1 + c(t4).w(t4,D77).1 = 0’02 x 1’16 + 0’43 x 1’52 =0’02+0’65= 0’67

sim(D84,Q) = c(t2).w(t2,D84).1 = 0’68 x 1’24 = 0’84

sim(D90,Q) = c(t3).w(t3,D90).1 = 0’02 x 1’53 = 0’03

sim(D93,Q) = c(t4).w(t4,D93).1 = 0’43 x 1’28 = 0’55

El resto de los 100 primeros documentos de la colección tienen un valor de similaridad 0 con la consulta Q, por lo que los eliminamos de la respuesta.

Por tanto, la respuesta del Sistema a Q sería:

D3 / D84 / D57 / D77 / D11 / D93 / D36 / D62 / D24 / D90 / D2

EJERCICIO 3

Sea un Sistema de Recuperación de Información cuya colección consta de los siguientes documentos:

D1: Una vez recopilados los datos y antes de someterlos al análisis, deben llevarse a cabo algunas operaciones preliminares, entre las que destacan la eliminación de datos incompletos o erróneos y la normalización.

D2: La eliminación de datos debe realizarse con precaución. No se trata de la eliminación de datos 'anómalos' frente al resto, sino de la eliminación de datos erróneos o irrelevantes.

D3: La normalización implica reducir los datos a una misma escala, en muchas ocasiones eliminando la influencia de algún factor conocido, pero sin interés en nuestro estudio.

D4: El análisis propiamente dicho de los datos trata de sacar a la luz un patrón o estructura a la que se ajustan dichos datos.

D5: Es habitual que el investigador tenga en mente, desde el inicio del estudio, un modelo o patrón para los datos recopilados, normalmente proporcionados por las hipótesis iniciales del estudio.

Las palabras subrayadas son los términos de indexación que tendrá en cuenta el sistema.

Considere los valores usuales para los parámetros k1, k3 y b:

$$k1 = 1'2;\ \ k3 = 0;\ \ b = 0'75$$

La longitud de los cinco documentos de la colección en los que aparecen los términos de la consulta es la siguiente:

D1 = 214 bytes; D2 = 174 bytes; D3 = 156 bytes; D4 = 119 bytes; D5 = 183 bytes;

Considere que la longitud media de los documentos de la colección es de 170 bytes:

$$avgdl = 170 \text{ bytes}$$

Tratándose del comienzo del proceso, considere que R = r = 0

Es suficiente utilizar dos decimales en los cálculos. A la hora de representar la respuesta, se indicará con el símbolo '/' el orden entre documentos, y una coma servirá para indicar que el sistema no puede imponer un orden entre documentos (debido a que poseen la misma similaridad con la consulta). Así, la respuesta D3 / D5 / D2, D4 / D1 indicaría que el sistema situaría en primer lugar, con la máxima similaridad, el documento D3; a continuación, con menor similaridad, el documento D5; a continuación, con la misma similaridad –aunque menor que la que posee el documento D5-, el sistema incluye los documentos D2 y D4; por último, con la menor similaridad, el documento D1.

Hallar la respuesta inicial del sistema ante la consulta:

$$Q = \text{patrón datos recopilados}$$

SOLUCIÓN

Hallamos los valores de la constante K para los documentos de la colección:

K(D1) = k1((1-b) + b.dl_{D1}/avgdl) = 1’2 ((1-0’75) + 0’75.214/170) = 1’43

K(D2) = k1((1-b) + b.dl_{D2}/avgdl) = 1’2 ((1-0’75) + 0’75.174/170) = 1’22

K(D3) = k1((1-b) + b.dl_{D3}/avgdl) = 1’2 ((1-0’75) + 0’75.156/170) = 1’13

K(D4) = k1((1-b) + b.dl_{D4}/avgdl) = 1’2 ((1-0’75) + 0’75.119/170) = 0’93

K(D5) = k1((1-b) + b.dl_{D5}/avgdl) = 1’2 ((1-0’75) + 0’75.183/170) = 1’27

Resumimos las frecuencias de los términos en los documentos en la siguiente tabla:

W_{td}

	D1	D2	D3	D4	D5
análisis	1			1	
anómalos		1			
datos	2	3	1	2	1
eliminación	1	3			
erróneos	1	1			
escala			1		
estructura				1	
estudio			1		2
factor			1		
hipótesis					1
incompletos	1				
irrelevantes		1			
modelo					1
normalización	1		1		
operaciones	1				
patrón				1	1
preliminares	1				
recopilados	1				1
reducir			1		

Resumimos los pesos de los términos de la consulta en los documentos en la siguiente tabla:

	D1	D2	D3	D4	D5
patrón	0	0	0	$\frac{2'2.1}{0'93+1}$	$\frac{2'2.1}{1'27+1}$
datos	$\frac{2'2.2}{1'43+2}$	$\frac{2'2.3}{1'22+3}$	$\frac{2'2.1}{1'12+1}$	$\frac{2'2.2}{0'93+2}$	$\frac{2'2.1}{1'27+1}$
recopilados	$\frac{2'2.1}{1'43+1}$	0	0	0	$\frac{2'2.1}{1'27+1}$

Una vez calculados los valores correspondientes a cada casilla, tenemos la siguiente tabla de pesos de los términos de la consulta en los documentos:

W_{td}

	D1	D2	D3	D4	D5
patrón	0	0	0	1’14	0’97
datos	0’64	1’56	1’04	1’50	0’97
recopilados	0’91	0	0	0	0’97

Resumimos los pesos de los términos en la consulta en la siguiente tabla:

W_{tq}

patrón	1
datos	1
recopilados	1

Calculamos a continuación los valores de los coeficientes correspondientes a los términos de la consulta:

c(patrón) = $\log\frac{5-2+0,5}{2+0,5}$ = log 1’4 = 0’15

c(datos) = $\log\frac{5-5+0,5}{5+0,5}$ = log 0’09 = -1’05

c(recopilados) = $\log\frac{5-2+0,5}{2+0,5}$ = log 1’4 = 0’15

Calculamos finalmente los valores de similaridad entre documentos y consulta:

sim(D1,Q) = c(datos).w(datos,D1).1 + c(recopilados).w(recopilados,D1).1 =

= -1’05 x 0’64 + 0’15 x 0’91 = -0’5355

sim(D2,Q) = c(datos).w(datos,D2).1 = -1’05 x 1’56 = -1’638

sim(D3,Q) = c(datos).w(datos,D3).1 = -1’05 x 1’04 = -1’1

sim(D4,Q) = c(datos).w(datos,D4).1 + c(patrón).w(patrón,D4).1= -1’05 x 1’5 + 0’15 x 1’14 =

= -1’40

sim(D5,Q) = c(datos).w(datos,D5) .1 + c(patrón).w(patrón,D5) .1 + c(recop).w(recop,D5) .1 =

= -1’05 x 0’97 + 0’15 x 0’97 + 0’15 x 0’97 = 0’15

Por tanto, la respuesta del Sistema a Q sería: D5 / D1 / D3 / D4 / D2

EJERCICIO 4

Sea un Sistema de Recuperación de Información compuesta por 3 millones de documentos. La matriz de ocurrencias término/documento es (solamente los cinco primeros documentos y los siguientes diez términos seleccionados):

	D1	D2	D3	D4	D5
científica	1	0	2	0	1
constante	0	0	1	0	0
crecimiento	0	2	1	1	0
exponencial	0	1	1	2	0
información	1	1	2	0	1
investigadores	0	2	0	1	0
lineal	3	0	0	1	0
literatura	1	0	1	0	2
Price	0	1	0	0	1
proporcional	1	0	1	0	0

Los términos de la consulta aparecen en los siguientes documentos de la colección:

crecimiento = 251040; exponencial = 517399; literatura = 1471863; científica = 806137

Considere los valores usuales para los parámetros k1, k3 y b:

$$k1 = 1'2;\ \ k3 = 0;\ \ b = 0'75$$

La longitud de los 5 primeros documentos de la colección es la siguiente:

D1 = 36 bytes; D2 = 32 bytes; D3 = 40 bytes; D4 = 30 bytes; D5 = 34 bytes

Considere que la longitud media de los documentos de la colección es de 30 bytes:

$$avgdl = 30 \text{ bytes}$$

Tratándose del comienzo del proceso, considere que R = r = 0

Es suficiente utilizar dos decimales en los cálculos. A la hora de representar la respuesta, se indicará con el símbolo '/' el orden entre documentos, y una coma servirá para indicar que el sistema no puede imponer un orden entre documentos (debido a que poseen la misma similaridad con la consulta). Así, la respuesta D3 / D5 / D2, D4 / D1 indicaría que el sistema situaría en primer lugar, con la máxima similaridad, el documento D3; a continuación, con menor similaridad, el documento D5; a continuación, con la misma similaridad –aunque menor que la que posee el documento D5-, el sistema incluye los documentos D2 y D4; por último, con la menor similaridad, el documento D1.

Hallar la respuesta inicial del sistema en relación a los cinco primeros documentos de la colección ante la consulta:

Q = crecimiento exponencial literatura científica

SOLUCIÓN

Hallamos los valores de la constante K para los cinco primeros documentos de la colección:

K(D1) = k1((1-b) + b.dl_{D1}/avgdl) = 1’2 ((1-0’75) + 0’75.36/30) = 1’38

K(D2) = k1((1-b) + b.dl_{D2}/avgdl) = 1’2 ((1-0’75) + 0’75.32/30) = 1’26

K(D3) = k1((1-b) + b.dl_{D3}/avgdl) = 1’2 ((1-0’75) + 0’75.40/30) = 1’5

K(D4) = k1((1-b) + b.dl_{D4}/avgdl) = 1’2 ((1-0’75) + 0’75.30/30) = 1’2

K(D5) = k1((1-b) + b.dl_{D5}/avgdl) = 1’2 ((1-0’75) + 0’75.34/30) = 1’32

Resumimos los pesos de los términos de la consulta en los documentos en la siguiente tabla:

W_{td}

	D1	D2	D3	D4	D5
crecimiento	0	$\frac{2'2.2}{1'26+2}$	$\frac{2'2.1}{1'5+1}$	$\frac{2'2.1}{1'2+1}$	0
exponencial	0	$\frac{2'2.1}{1'26+1}$	$\frac{2'2.1}{1'5+1}$	$\frac{2'2.2}{1'2+2}$	0
literatura	$\frac{2'2.1}{1'38+1}$	0	$\frac{2'2.1}{1'5+1}$	0	$\frac{2'2.2}{1'32+2}$
científica	$\frac{2'2.1}{1'38+1}$	0	$\frac{2'2.2}{1'5+2}$	0	$\frac{2'2.1}{1'32+1}$

Una vez calculados los valores correspondientes a cada casilla, tenemos la siguiente tabla de pesos de los términos de la consulta en los documentos:

W_{td}

	D1	D2	D3	D4	D5
crecimiento	0	1’35	0’88	1	0
exponencial	0	0’97	0’88	1’38	0
literatura	0’92	0	0’88	0	1’33
científica	0’92	0	1’26	0	0’95

Resumimos los pesos de los términos en la consulta en la siguiente tabla:

W_{tq}

crecimiento	1
exponencial	1
literatura	1
científica	1

Calculamos a continuación los valores de los coeficientes correspondientes a los términos de la consulta:

c(crecimiento) = $\log\frac{3000000-251040+0,5}{251040+0,5}$ = log 10’95 = 1’04

c(exponencial) = $\log\frac{3000000-517399+0,5}{517399+0,5}$ = log 4’80 = 0’68

c(literatura) = $\log\frac{3000000-1471863+0,5}{1471863+0,5}$ = log 1’04 = 0’02

c(científica) = $\log\frac{3000000-806137+0,5}{806137+0,5}$ = log 2’72 = 0’43

Calculamos finalmente los valores de similaridad entre documentos y consulta:

sim(D1,Q) = c(literatura).w(literatura,D1).1 + c(científica).w(científica,D1).1 =

= 0’02 x 0’92 + 0’43 x 0’92 = 0’02 + 0’40 = 0’42

sim(D2,Q) = c(crecimiento).w(crecimiento,D2).1 + c(exponencial).w(exponencial,D2).1 =

= 1’04 x 1’35 + 0’68 x 0’97 = 1’40 + 0’66 = 2’06

sim(D3,Q) = c(crecimiento).w(crecimiento,D3).1 + c(exponencial).w(exponencial,D3).1 +

+ c(literatura).w(literatura,D3).1 + c(científica).w(científica,D3).1 =

= 1’04 x 0’88 + 0’68 x 0’88 + 0’02 x 0’88 + 0’43 x 1’26 = 0’92 + 0’60 + 0’02 + 0’54 = 2’08

sim(D4,Q) = c(crecimiento).w(crecimiento,D4).1 + c(exponencial).w(exponencial,D4).1=

= 1’04 x 1 + 0’68 x 1’38 = 1’04 + 0.94 = 1’98

sim(D5,Q) = c(literatura).w(literatura,D5) .1 + c(científica).w(científica,D5) .1 =

= 0’02 x 1’33 + 0’43 x 0’95 = 0’03 + 0’41 = 0’44

Por tanto, la respuesta del Sistema a Q sería: D3 / D2 / D4 / D5 / D1

EJERCICIO 5

Sea un Sistema de Recuperación de Información cuya colección consta de 2 millones de documentos. La información relativa a los primeros cuatro documentos de dicha colección (los términos de indexación que contienen, la frecuencia de aparición y sus posiciones) se resumen en la siguiente tabla:

	TÉRMINOS (FRECUENCIA, <POSICIONES>)
D1	bibliometría(1,<44>); cienciometría(1,<21>); método(1,<108>); matemático(2,<6,41>); estadístico(1,<32>); literatura(1,<57>); científica(1,<89>)
D2	producción(1,<24>); análisis(2,<13,48>); actividad(1,<22>); científica(1,<39>); leyes(2,<90,231>); bibliometría(2,<4,11>)
D3	leyes(1,<83>); comportamiento(1,<31>); estadístico(2,<2,16>); regular(1,<25>); ciencia(1,<64>)
D4	medición(1,<213>); estadístico(1,<118>); indicador(1,<40>); bibliometría(1,<105>); actividad(2,<190,202>); científica(2,<211,374)

Los términos de la consulta aparecen en los siguientes documentos de la colección:

leyes = 125520; ciencia = 258695; bibliometría = 735931

Considere los valores usuales para los parámetros k1, k3 y b:

$$k1 = 1'2;\ \ k3 = 0;\ \ b = 0'75$$

La longitud de los 4 primeros documentos de la colección es la siguiente:

D1 = 110 bytes; D2 = 250 bytes; D3 = 90 bytes; D4 = 380 bytes

Considere que la longitud media de los documentos de la colección es de 280 bytes:

$$avgdl = 280 \text{ bytes}$$

Tratándose del comienzo del proceso, considere que R = r = 0

Es suficiente utilizar dos decimales en los cálculos. A la hora de representar la respuesta, se indicará con el símbolo '/' el orden entre documentos, y una coma servirá para indicar que el sistema no puede imponer un orden entre documentos (debido a que poseen la misma similaridad con la consulta). Así, la respuesta D3 / D5 / D2, D4 / D1 indicaría que el sistema situaría en primer lugar, con la máxima similaridad, el documento D3; a continuación, con menor similaridad, el documento D5; a continuación, con la misma similaridad –aunque menor que la que posee el documento D5-, el sistema incluye los documentos D2 y D4; por último, con la menor similaridad, el documento D1.

Hallar la respuesta inicial del sistema en relación a los cuatro primeros documentos de la colección ante la consulta:

$$Q = \text{leyes ciencia bibliometria}$$

SOLUCIÓN

Hallamos los valores de la constante K para los 4 primeros documentos de la colección:

K(D1) = k1((1-b) + b.dl_{D1}/avgdl) = 1’2 ((1-0’75) + 0’75.110/280) = 0’65

K(D2) = k1((1-b) + b.dl_{D2}/avgdl) = 1’2 ((1-0’75) + 0’75.250/280) = 1’1

K(D3) = k1((1-b) + b.dl_{D3}/avgdl) = 1’2 ((1-0’75) + 0’75.90/280) = 0’59

K(D4) = k1((1-b) + b.dl_{D4}/avgdl) = 1’2 ((1-0’75) + 0’75.380/280) = 1’5

Resumimos los pesos de los términos de la consulta en estos documentos en la siguiente tabla:

W_{td}

	D1	D2	D3	D4
leyes	0	$\frac{2'2.2}{1'1+2}$	$\frac{2'2.1}{0'59+1}$	0
ciencia	0	0	$\frac{2'2.1}{0'59+1}$	0
bibliometría	$\frac{2'2.1}{0'65+1}$	$\frac{2'2.2}{1'1+2}$	0	$\frac{2'2.1}{1'5+1}$

Una vez calculados los valores correspondientes a cada casilla, tenemos:

W_{td}

	D1	D2	D3	D4
leyes	0	1’42	1’38	0
ciencia	0	0	1’38	0
bibliometría	1’33	1’42	0	0’88

Resumimos los pesos de los términos en la consulta en la siguiente tabla:

W_{tq}

leyes	1
ciencia	1
bibliometría	1

Calculamos a continuación los valores de los coeficientes correspondientes a los términos de la consulta:

c(leyes) = $\log\frac{2000000-125520+0,5}{125520+0,5}$ = log 14’93 = 1’17

c(ciencia) = $\log\frac{2000000-258695+0,5}{258695+0,5}$ = log 6’73 = 0’83

c(bibliometría) = $\log\frac{2000000-735931+0,5}{735931+0,5}$ = log 1’72 = 0’24

Calculamos finalmente los valores de similaridad entre documentos y consulta:

sim(D1,Q) = c(bibliometría).w(bibliometría,D1).1 = 0’24 x 1’33 = 0’32

sim(D2,Q) = c(leyes).w(leyes,D2) + c(bibliometría).w(bibliometría,D2) =

= 1’17 x 1’42 + 0’24 x 1’42 = 0’49 + 0’34 = 0’83

sim(D3,Q) = c(leyes).w(leyes,D3).1 + c(ciencia).w(ciencia,D3).1 = 1’17 x 1’38 + 0’83 x 1’38 =

= 1’61 + 1’15 = 2’76

sim(D4,Q) = c(bibliometría).w(bibliometría,D4).1 = 0’24 x 0’88 = 0’21

Por tanto, la respuesta del Sistema a Q en relación a los cuatro primeros documentos de la colección sería:

D3 / D2 / D1 / D4

EJERCICIO 6

Sea un Sistema de Recuperación de Información cuya colección consta de 5 millones de documentos. La colección incluye 672000 términos de indexación, de los cuales los términos incluidos en las últimas consultas de un usuario tienen la siguiente distribución:

	N
T1	636199
T2	218775
T3	762903
T4	1286432
T5	1043843

Siendo 'n' el número de documentos de la colección en los que aparece el término correspondiente. Cinco de los documentos de la colección están representados de la siguiente manera, en relación a los términos antes señalados:

	T1	T2	T3	T4	T5
D2	1	0	0	2	0
D908	0	1	0	0	1
D1001	2	0	0	1	0
D356411	0	2	1	0	1
D703246	1	1	0	0	0

Considere los valores usuales para los parámetros k1, k3 y b:

$$k1 = 1'2;\ \ k3 = 0;\ \ b = 0'75$$

La longitud de los citados documentos de la colección es la siguiente:

D2 = 47 bytes; D908 = 39 bytes; D1001 = 41 bytes; D356411 = 62 bytes; D703246 = 36 bytes;

Considere que la longitud media de los documentos de la colección es de 50 bytes:

$$avgdl = 50 \text{ bytes}$$

Tratándose del comienzo del proceso, considere que R = r = 0

Es suficiente utilizar dos decimales en los cálculos. A la hora de representar la respuesta, se indicará con el símbolo '/' el orden entre documentos, y una coma servirá para indicar que el sistema no puede imponer un orden entre documentos (debido a que poseen la misma similaridad con la consulta). Así, la respuesta D3 / D5 / D2, D4 / D1 indicaría que el sistema situaría en primer lugar, con la máxima similaridad, el documento D3; a continuación, con menor similaridad, el documento D5; a continuación, con la misma similaridad –aunque menor que la que posee el documento D5-, el sistema incluye los documentos D2 y D4; por último, con la menor similaridad, el documento D1.

Hallar la respuesta inicial del sistema en relación a los citados documentos de la colección ante la consulta: Q = T3 T5 T1

SOLUCIÓN

Hallamos los valores de la constante K para los documentos de la colección señalados:

K(D2) = k1((1-b) + b.dl_{D1}/avgdl) = 1’2 ((1-0’75) + 0’75.47/50) = 1’15

K(D908) = k1((1-b) + b.dl_{D2}/avgdl) = 1’2 ((1-0’75) + 0’75.39/50) = 1’00

K(D1001) = k1((1-b) + b.dl_{D3}/avgdl) = 1’2 ((1-0’75) + 0’75.41/50) = 1’04

K(D356411) = k1((1-b) + b.dl_{D4}/avgdl) = 1’2 ((1-0’75) + 0’75.62/50) = 1’42

K(D703246) = k1((1-b) + b.dl_{D5}/avgdl) = 1’2 ((1-0’75) + 0’75.36/50) = 0’95

Resumimos los pesos de los términos de la consulta en estos documentos en la siguiente tabla:

W_{td}

	D2	D908	D1001	D356411	D703246
T1	$\frac{2'2.1}{1'15+1}$	0	$\frac{2'2.2}{1'04+2}$	0	$\frac{2'2.1}{0'95+1}$
T3	0	0	0	$\frac{2'2.1}{1'42+1}$	0
T5	0	$\frac{2'2.1}{1+1}$	0	$\frac{2'2.1}{1'42+1}$	0

Una vez calculados los valores correspondientes a cada casilla, tenemos:

W_{td}

	D2	D908	D1001	D356411	D703246
T1	1’02	0	1’45	0	1’13
T3	0	0	0	0’91	0
T5	0	1’1	0	0’91	0

Resumimos los pesos de los términos en la consulta en la siguiente tabla:

W_{tq}

T1	1
T3	1
T5	1

Calculamos a continuación los valores de los coeficientes correspondientes a los términos de la consulta:

c(T1) = $\log\frac{5000000-636199+0,5}{636199+0,5}$ = log 6’86 = 0’84

c(T3) = $\log\frac{5000000-762903+0,5}{762903+0,5}$ = log 5’55 = 0’74

c(T5) = $\log\frac{5000000-1043843+0,5}{1043843+0,5}$ = log 3’79 = 0’58

Calculamos finalmente los valores de similaridad entre documentos y consulta:

sim(D2,Q) = c(T1).w(T1,D2).1 = 0’84 x 1’02 = 0’86

sim(D908,Q) = c(T5).w(T5,D908).1= 0’58 x 1’1 = 0’64

sim(D1001,Q) = c(T1).w(T1,D1001).1 = 0’84 x 1’45 = 1’22

sim(D356411,Q) = c(T3).w(T3,D356411).1 + c(T5).w(T5,D356411).1 = 0’74 x 0’91 + 0’58 x 0’91=

= 0’67 + 0’53 = 1’20

sim(D703246,Q) = c(T1).w(T1,D703246).1 = 0’84 x 1’13 = 0’95

Por tanto, la respuesta del Sistema a Q en relación a los cuatro primeros documentos de la colección sería:

D1001 / D356411 / D703246 / D2 / D908

MODELO VECTORIAL CON NORMALIZACIÓN POR LONGITUD BASADA EN PIVOTE

Este modelo vectorial con normalización por longitud basada en pivote sigue los mismos principios del modelo vectorial clásico, pero es consciente de que en el modelo clásico los documentos largos tienen más posibilidades de ser recuperados como relevantes que los documentos cortos ante cualquier consulta. En efecto, cuanto más largo es el documento más tiende a aumentar la frecuencia de los términos presentes en el mismo; además, al contener más términos, hay mayor probabilidad de que incluya un número superior de términos de entre los que se hayan introducido en las consultas.

Para paliar el sesgo en la respuesta del sistema con la inclusión primordialmente de documentos largos frente a documentos cortos, este modelo añade un factor de normalización, de manera que las probabilidades de aparición de un documento como relevante en la respuesta no dependa de su longitud. Esta normalización suele constar de tres modificaciones complementarias:

1. El valor clásico del IDF penaliza en exceso los términos corrientes que aparecen en un número grande de documentos de la colección, aunque no aparezcan en un número excesivo o siquiera mayoritario. Para suavizar este comportamiento, de manera que los términos tengan un IDF muy semejante entre sí en cuanto aparezcan en más del 50% de los documentos de la colección, se adopta la siguiente fórmula para el cálculo del IDF de un término:

$$IDF\ (término\ t) = log\frac{(N+1)}{n}$$

Donde:

- N es el número de documentos de la colección
- n es el número de documentos de la colección que contienen el término 't'

2. El valor de TF impuesto por el modelo vectorial clásico premia, como hemos dicho previamente, a los documentos largos, pues los términos tienden a aparecer con mayor frecuencia cuanto más largo es el documento. Para suavizar este comportamiento, de manera que la diferencia entre los valores de TF altos y bajos sea menor que si imponemos una variación lineal, se suele adoptar la siguiente fórmula para el cálculo del TF de un término en un documento:

$$TF\ (término\ t\ en\ documento\ d) = 1 + \log(1 + \log(tf_{td}))$$

Donde:

- tf_{td} es el número de veces que aparece el término 't' en el documento 'd'

De esta forma, la fórmula del pesado de los términos en los documentos de la colección posee ahora la siguiente expresión:

$$w_{td} = TF_{td} \,.\, IDF_t = [1 + \log(1 + \log(tf_{td}))].\, log\, \frac{(N+1)}{n}$$

Donde:

- w_{td} es el peso correspondiente al término 't' en el documento 'd'
- TF_{td} es el valor de TF correspondiente al término 't' en el documento 'd'
- IDF_t es el valor de IDF correspondiente al término 't'
- tf_{td} es el número de veces que aparece el término 't' en el documento 'd'
- N es el número de documentos de la colección
- n es el número de documentos de la colección que contienen el término 't'

Sin embargo, la fórmula habitualmente empleada para el cálculo de los pesos de los términos en las consultas no es distinta de la empleada en el modelo vectorial clásico, sobreentendiendo que la longitud de las consultas es siempre mucho más corta que la longitud de los documentos, de manera que no requiere ningún factor de corrección. Por tanto, si el cálculo del peso de los términos en la consulta se basa exclusivamente en la frecuencia de aparición de dichos términos en la misma y no en una cantidad numérica impuesta directa o indirectamente por el usuario, su expresión es simplemente:

$$w_{tq} = \, tf_{tq}$$

Donde:

- w_{tq} es el peso correspondiente al término 't' en la consulta 'q'
- tf_{tq} es la frecuencia de aparición del término 't' en la consulta 'q'

3. Además, a la hora de calcular la similaridad entre un documento 'd' y una consulta 'q', se introduce un factor de normalización al cálculo realizado en el modelo clásico, esto es, el producto escalar o producto punto de los pesos correspondientes a los términos presentes simultáneamente en el documento 'd' y la consulta 'q' (recordemos que el producto escalar o producto punto se efectúa en dos pasos: 1) se multiplican primero cada uno de los pesos de los términos presentes en el documento 'd' por los pesos de

cada uno de esos mismos términos en la consulta ‘q’; 2) posteriormente se suman estas cantidades, cada una relativa a un término común entre el documento ‘d’ y la consulta ‘q’). Como decimos, el factor de normalización implica la inclusión en el cálculo de la similaridad de un parámetro ‘s’ que depende de la colección concreta, aunque actualmente s= 0’2 suele ser el valor más recomendado. De igual forma, este factor de normalización incluye también la longitud en bytes del documento (dl_d) y la longitud media en bytes de los documentos de la colección (avgdl). La fórmula de la similaridad queda finalmente:

$$Sim(d,q) = \sum_{t \in d\ y\ q} \frac{1}{(1-s) + s.\frac{dl_d}{avgdl}} . w_{td} . w_{tq}$$

Donde:

- Sim(d,q) es el valor de similaridad entre el documento ‘d’ y la consulta ‘q’
- w_{td} es el peso correspondiente al término ‘t’ en el documento ‘d’
- w_{tq} es el peso correspondiente al término ‘t’ en la consulta ‘q’
- ‘s’ es el parámetro de normalización (normalmente s = 0’2)
- ‘dl_d’ es la longitud en bytes del documento ‘d’
- ‘avgdl’ es la longitud media en bytes de los documentos de la colección

En conclusión, la fórmula final de la similaridad entre un documento ‘d’ y una consulta ‘q’, cuando el peso de los términos en la consulta se calcula considerando exclusivamente la frecuencia de aparición, es la siguiente:

$$Sim(d,q) = \sum_{t \in d\ y\ q} \frac{1}{(1-s) + s.\frac{dl_d}{avgdl}} . [1 + \log(1 + \log(tf_{td}))] . \ log\frac{(N+1)}{n} . \ tf_{tq}$$

Donde:

- Sim(d,q) es el valor de similaridad entre el documento ‘d’ y la consulta ‘q’
- ‘s’ es el parámetro de normalización (normalmente s = 0’2)
- ‘dl_d’ es la longitud en bytes del documento ‘d’
- ‘avgdl’ es la longitud media en bytes de los documentos de la colección
- tf_{td} es la frecuencia de aparición del término ‘t’ en el documento ‘d’
- N es el número de documentos de la colección
- n es el número de documentos de la colección que contienen el término ‘t’
- tf_{tq} es la frecuencia de aparición del término ‘t’ en la consulta ‘q’

Esta fórmula de la similaridad muestra la semejanza del modelo probabilístico BM25 con los modelos vectoriales que efectúan normalización de la longitud de los documentos. Puede comprobarse dicha similitud consultando, por ejemplo, la fórmula de la similaridad en el modelo probabilístico BM25 (previamente, en este mismo documento).

BIBLIOGRAFÍA

Losada, David E. Modelos de recuperación de información II. En recuperación de información: Un enfoque práctico y multidisciplinar. Madrid: RA-MA, 2011, pp. 295-358.

Singhal, A. Modern information retrieval: a brief overview. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, vol. 24(4), pp. 35-43.

Singhal, A. Buckley, C.; Mitra, M. Pivoted document length normalization. En Proceedings of the 19th ACM SIGIR Conference. New York: ACM, 1996, pp. 21-29.

EJERCICIO 1

Sea un Sistema de Recuperación de Información cuya colección consta de 4 millones de documentos. Las consultas realizadas por el último usuario incluían los términos 'análisis', 'documental', 'descripción', 'física' y 'formal'. Los datos estadísticos esenciales en relación a dichos términos se resumen en la siguiente tabla:

	N	LISTA
análisis	271684	(19,1),(27,2),(54,2),(63,2),(71,1)
documental	150542	(19,1),(38,1),(54,2),(72,1)
descripción	164326	(11,2),(24,1),(83,1),(99,1)
física	133896	(54,1),(63,1),(99,2)
formal	26098	(11,1),(90,1)

Donde 'n' es el número de documentos de la colección en los que aparece el término. La 'Lista' incluye los números de los 100 primeros documentos de la colección que contienen cada término, así como la frecuencia de aparición del término en cada documento. Así, (79,3) indica que el término aparece 3 veces en el documento número 79.

Considere el valor habitual para el parámetro 's':

$$s = 0'2$$

La longitud de los 100 primeros documentos de la colección en los que aparecen los términos de la consulta es la siguiente:

D11 = 36 bytes; D19 = 28 bytes; D24 = 42 bytes; D38 = 50 bytes; D54 = 44 bytes;

D63 = 30 bytes; D72 = 32 bytes; D83 = 48 bytes; D99 = 46 bytes

Considere que la longitud media de los documentos de la colección es de 45 bytes:

$$avgdl = 45 \text{ bytes}$$

Calcule el peso de los términos en la consulta basándose exclusivamente en la frecuencia de aparición de dichos términos en la misma.

Es suficiente utilizar dos decimales en los cálculos. A la hora de representar la respuesta, se indicará con el símbolo '/' el orden entre documentos, y una coma servirá para indicar que el sistema no puede imponer un orden entre documentos (debido a que poseen la misma similaridad con la consulta). Así, la respuesta D3 / D5 / D2, D4 / D1 indicaría que el sistema situaría en primer lugar, con la máxima similaridad, el documento D3; a continuación, con menor similaridad, el documento D5; a continuación, con la misma similaridad –aunque menor que la que posee el documento D5-, el sistema incluye los documentos D2 y D4; por último, con la menor similaridad, el documento D1.

Hallar la respuesta inicial del sistema en relación a los 100 primeros documentos de la colección ante la consulta: Q = descripción física documental

SOLUCIÓN

Resumimos la matriz de ocurrencias término/documento en la siguiente tabla, incluyendo solamente los términos empleados en la consulta:

tf_{td} (número de veces que aparece un término en un documento)[casilla vacía implica un cero]

	DESCRIPCIÓN	FÍSICA	DOCUMENTAL
D11	2		
D19			1
D24	1		
D38			1
D54		1	2
D63		1	
D72			1
D83	1		
D99	1	2	

Ahora calculamos los valores de $TF_{td} = 1 + \log(1 + \log(tf_{td}))$ [casilla vacía implica un cero]

	DESCRIPCIÓN	FÍSICA	DOCUMENTAL
D11	1’11		
D19			1
D24	1		
D38			1
D54		1	1’11
D63		1	
D72			1
D83	1		
D99	1	1’11	

Los valores de IDF de cada uno de los términos empleados en la consulta son los siguientes:

IDF (descripción) = $\log \frac{4000000+1}{164326}$ = log 24’34 = 1’39

IDF (física) = $\log \frac{4000000+1}{133896}$ = log 29’87 = 1’48

IDF (documental) = $\log \frac{4000000+1}{150542}$ = log 26’57 = 1’42

Los valores de los pesos de los términos en los documentos se resumen en la siguiente tabla (casilla vacía implica un cero):

W_{td}

	DESCRIPCIÓN	FÍSICA	DOCUMENTAL
D11	1’54		
D19			1’42
D24	1’39		
D38			1’42
D54		1’48	1’58
D63		1’48	
D72			1’42
D83	1’39		
D99	1’39	1’64	

El peso de los términos en la consulta, basándose exclusivamente en la frecuencia de aparición, es igual a la unidad:

W_{tq}

	DESCRIPCIÓN	FÍSICA	DOCUMENTAL
Q	1	1	1

El cálculo de los parámetros de normalización ‘K’ correspondientes a cada documento en el que aparece alguno de los términos de la consulta da los siguientes valores:

K(D11) = $\frac{1}{(1-0,2)+0,2.\frac{36}{45}} = \frac{1}{0,96}$ = 1’04

K(D19) = $\frac{1}{(1-0,2)+0,2.\frac{28}{45}} = \frac{1}{0,92}$ = 1’09

K(D24) = $\frac{1}{(1-0,2)+0,2.\frac{42}{45}} = \frac{1}{0,99}$ = 1’01

K(D38) = $\frac{1}{(1-0,2)+0,2.\frac{50}{45}} = \frac{1}{1,02}$ = 0’98

K(D54) = $\frac{1}{(1-0,2)+0,2.\frac{44}{45}} = \frac{1}{1}$ = 1

K(D63) = $\frac{1}{(1-0,2)+0,2.\frac{30}{45}} = \frac{1}{0,93}$ = 1’08

K(D72) = $\frac{1}{(1-0,2)+0,2.\frac{32}{45}} = \frac{1}{0,94}$ = 1’06

K(83) = $\frac{1}{(1-0,2)+0,2.\frac{48}{45}} = \frac{1}{1,01}$ = 0’99

K(D99) = $\frac{1}{(1-0,2)+0,2.\frac{46}{45}} = \frac{1}{1}$ = 1

Finalmente, el cálculo de similaridades entre los documentos de la colección y la consulta es el siguiente:

Sim(D11,Q) = K(D11). w(descripción,D11).w(descripción,Q) = 1’04.1’54.1 = 1’60

Sim(D19,Q) = K(D19). w(documental,D19).w(documental,Q) = 1’09.1’42.1 = 1’55

Sim(D24,Q) = K(D24). w(descripción,D24).w(descripción,Q) = 1’01.1’39.1 = 1’40

Sim(D38,Q) = K(D38). w(documental,D38).w(documental,Q) = 0’98.1’42 = 1’39

Sim(D54,Q) = K(D54). w(física,D54).w(física,Q) + K(D54). w(documental,D54).w(documental,Q)

= 1.1’48.1 + 1.1’58.1 = 3’06

Sim(D63,Q) = K(D63). w(física,D63).w(física,Q) = 1’08.1’48.1 = 1’60

Sim(D72,Q) = K(D72). w(documental,D72).w(documental,Q) = 1’06.1’42.1 = 1’51

Sim(D83,Q) = K(D83). w(descripción,D83).w(descripción,Q) = 0’99.1’39.1 = 1’38

Sim(D99,Q) = K(D99). w(descripción,D99).w(descripción,Q) + K(D99). w(física,D99).w(física,Q)

= 1.1’39.1 + 1.1’64.1 = 3’03

El resto de los 100 primeros documentos de la colección tienen un valor de similaridad 0 con la consulta Q, por lo que los eliminamos de la respuesta.

Por tanto, la respuesta del Sistema a Q sería la siguiente:

D54 / D99 / D11, D63 / D19 / D72 / D24 / D38 / D83

EJERCICIO 2

Sea un Sistema de Recuperación de Información cuya colección consta de 5 millones de documentos. Los datos estadísticos esenciales en relación a los principales términos de indexación empleados en las últimas consultas se resumen en la siguiente tabla:

	N	LISTA DE DOCUMENTOS (LOS 200 PRIMEROS)	LISTA DE FRECUENCIA Y POSICIONES
T1	485298	127,157,183	(127,1,<45>),(157,2,<5,67>),(183,1,<224>)
T2	2943726	2,3,89,127,183	(2,1,<9>),(3,1,<542>),(89,1,<101>),(127,2,<11,534>), (183,1,<231>)
T3	130846	7,25	(7,2,<43,64>),(25,2,<66,74>)
T4	1034798	89,139,143,171	(89,2,<120,243>),(139,1,<7>),(143,1,<86>), (171,2,<8,13>)
T5	1612274	25,75,127,143	(25,1,<153>),(75,2,<57,201>),(127,2,<10,206>), (143,1,<235>)

Donde 'n' es el número de documentos de la colección en los que aparece el término. La lista de documentos incluye los números de los 200 primeros documentos de la colección que contienen cada término. La lista de frecuencias y posiciones incluye el número de documento donde surge el término, la frecuencia de aparición en dicho documento y sus posiciones absolutas. Así, (79,1,<40>) indica que el término aparece en el documento número 79, una sola vez, en la posición 40 del documento.

Considere el valor habitual para el parámetro 's':

$$s = 0'2$$

La longitud de los 200 primeros documentos de la colección en los que aparecen los términos de la consulta es la siguiente:

D7 = 159 bytes; D25 = 549 bytes; D75 = 265 bytes; D127 = 658 bytes; D143 = 444 bytes;

D157 = 603 bytes; D183 = 527 bytes

Considere que la longitud media de los documentos de la colección es de 545 bytes:

$$avgdl = 545 \text{ bytes}$$

Calcule el peso de los términos en la consulta como si se tratase de un documento más, empleando el IDF del término en la colección. Los números entre paréntesis señalan la frecuencia del término correspondiente en la consulta. Así, información(3) indicaría que el término 'información' aparece 3 veces en la consulta.

Es suficiente utilizar dos decimales en los cálculos. A la hora de representar la respuesta, se indicará con el símbolo '/' el orden entre documentos, y una coma servirá para indicar que el sistema no puede imponer un orden entre documentos (debido a que poseen la misma similaridad con la consulta). Así, la respuesta D3 / D5 / D2, D4 / D1 indicaría que el sistema

situaría en primer lugar, con la máxima similaridad, el documento D3; a continuación, con menor similaridad, el documento D5; a continuación, con la misma similaridad –aunque menor que la que posee el documento D5-, el sistema incluye los documentos D2 y D4; por último, con la menor similaridad, el documento D1.

Hallar la respuesta inicial del sistema en relación a los 200 primeros documentos de la colección ante la consulta:

$$Q = T1(1)\ \ T3(2)\ \ T5(1)$$

SOLUCIÓN

Resumimos la matriz de ocurrencias término/documento en la siguiente tabla, incluyendo solamente los términos empleados en la consulta:

tf_{td} (número de veces que aparece un término en un documento)[casilla vacía implica un cero]

	T1	T3	T5
D7		2	
D25		2	1
D75			2
D127	1		2
D143			1
D157	2		
D183	1		

Ahora calculamos los valores de $TF_{td} = 1 + \log(1 + \log(tf_{td}))$ [casilla vacía implica un cero]

	T1	T3	T5
D7		1’11	
D25		1’11	1
D75			1’11
D127	1		1’11
D143			1
D157	1’11		
D183	1		

Los valores de IDF de cada uno de los términos empleados en la consulta son los siguientes:

IDF (T1) = $\log \frac{5000000+1}{485298}$ = log 10’30 = 1’01

IDF (T3) = $\log \frac{5000000+1}{130846}$ = log 38’21 = 1’58

IDF (T5) = $\log \frac{5000000+1}{1612274}$ = log 3’10 = 0’49

Los valores de los pesos de los términos en los documentos se resumen en la siguiente tabla (casilla vacía implica un cero):

W_{td}

	T1	T3	T5
D7		1’75	
D25		1’75	0’49
D75			0’54
D127	1’01		0’54
D143			0’49
D157	1’12		
D183	1’01		

Para calcular el peso de los términos en la consulta, consideramos inicialmente su frecuencia de aparición:

tf_{tq} (número de veces que aparece un término en un documento)

	T1	T3	T5
Q	1	2	1

Al considerar la consulta como un documento más, calculamos a continuación los valores de $TF_{tq} = 1 + \log(1 + \log(tf_{tq}))$

	T1	T3	T5
Q	1	1’11	1

Con estos valores, calculamos los pesos de los términos en la consulta empleando los IDF de los términos en la colección:

$W_{tq} = TF_{tq} \cdot IDF_t$

	T1	T3	T5
Q	1’01	1’75	0’49

El cálculo de los parámetros de normalización ‘K’ correspondientes a cada documento en el que aparece alguno de los términos de la consulta da los siguientes valores:

$K(D7) = \frac{1}{(1-0,2)+0,2.\frac{159}{545}} = \frac{1}{0,86} = 1’16$

$K(D25) = \frac{1}{(1-0,2)+0,2.\frac{549}{545}} = \frac{1}{1} = 1$

$K(D75) = \frac{1}{(1-0,2)+0,2.\frac{265}{545}} = \frac{1}{0,9} = 1’11$

$K(D127) = \frac{1}{(1-0,2)+0,2.\frac{658}{545}} = \frac{1}{1,04} = 0’96$

K(D143) = $\frac{1}{(1-0,2)+0,2.\frac{444}{545}} = \frac{1}{0,96}$ = 1’04

K(D157) = $\frac{1}{(1-0,2)+0,2.\frac{603}{545}} = \frac{1}{1,02}$ = 0’98

K(D183) = $\frac{1}{(1-0,2)+0,2.\frac{527}{545}} = \frac{1}{0,99}$ = 1’01

Finalmente, el cálculo de similaridades entre los documentos de la colección y la consulta es el siguiente:

Sim(D7,Q) = K(D7). w(T3,D7).w(T3,Q) = 1’16 . 1’75 . 1’75 = 3’55

Sim(D25,Q) = K(D25). w(T3,D25).w(T3,Q) + K(D25). w(T5,D25).w(T5,Q) =

= 1 . 1’75 . 1’75 + 1. 0’49 . 0’49 = 3’06 + 0’24 = 3’3

Sim(D75,Q) = K(D75). w(T5,D75).w(T5,Q) = 1’11 . 0’54 . 0’49 = 0’29

Sim(D127,Q) = K(D127). w(T1,D127).w(T1,Q) + K(D127). w(T5,D127).w(T5,Q)=

= 0’96 . 1’01 . 1’01 + 0’96 . 0’54 . 0’49 = 0’98 + 0’25 = 1’23

Sim(D143,Q) = K(D143). w(T5,D143).w(T5,Q) = 1’04 . 0’49 . 0’49 = 0’25

Sim(D157,Q) = K(D157). w(T1,D157).w(T1,Q) = 0’98 . 1’12 . 1’01 = 1’11

Sim(D183,Q) = K(D183). w(T1,D183).w(T1,Q) = 1’01 . 1’01 . 1’01 = 1’03

El resto de los 200 primeros documentos de la colección tienen un valor de similaridad 0 con la consulta Q, por lo que los eliminamos de la respuesta.

Por tanto, la respuesta del Sistema a Q sería la siguiente:

D7 / D25 / D127 / D157 / D183 / D75 / D143

EJERCICIO 3

Sea un Sistema de Recuperación de Información cuya colección consta de los siguientes documentos:

D1: La irrupción de las bibliotecas digitales ha provocado la proliferación de servicios de digitalización de documentos en los formatos más diversos.

D2: El auge de las bibliotecas digitales ha conllevado un aumento en los estudios sobre su gestión, con el fin de unificar criterios y de proporcionar a los usuarios un servicio lo más eficaz y completo posible.

D3: Conforme a las recomendaciones de la Unión Europea, los organismos públicos deben promover la creación de bibliotecas digitales, impulsando la digitalización de colecciones analógicas y la accesibilidad en línea del patrimonio documental.

D4: El nivel más avanzado consiste en la aportación de sus colecciones digitales a Europeana. Para ello el repositorio OAI debe servir los registros al menos en formato ESE (Europeana Semantic Elements) o en EDM (Europeana Data Model) y, además, firmar el DEA (Data Exchange Agreement) con Europeana.

Las palabras subrayadas son los términos de indexación que tendrá en cuenta el sistema. Considere el valor habitual para el parámetro 's':

$$s = 0'2$$

La longitud de los documentos de la colección es la siguiente:

D1 = 167 bytes; D2 = 238 bytes; D3 = 306 bytes; D4 = 394 bytes

Considere que la longitud media de los documentos de la colección es de 245 bytes:

$$avgdl = 245 \text{ bytes}$$

Calcule el peso de los términos en la consulta basándose exclusivamente en la frecuencia de aparición de dichos términos en la misma. Así, información(3) indicaría que el término 'información' aparece 3 veces en la consulta.

Es suficiente utilizar dos decimales en los cálculos. A la hora de representar la respuesta, se indicará con el símbolo '/' el orden entre documentos, y una coma servirá para indicar que el sistema no puede imponer un orden entre documentos (debido a que poseen la misma similaridad con la consulta). Así, la respuesta D3 / D5 / D2, D4 / D1 indicaría que el sistema situaría en primer lugar, con la máxima similaridad, el documento D3; a continuación, con menor similaridad, el documento D5; a continuación, con la misma similaridad –aunque menor que la que posee el documento D5-, el sistema incluye los documentos D2 y D4; por último, con la menor similaridad, el documento D1.

Hallar la respuesta inicial del sistema en relación a la consulta:

Q = digitalización(3) documentos(1) bibliotecas(2)

SOLUCIÓN

Resumimos la matriz de ocurrencias término/documento en la siguiente tabla, incluyendo solamente los términos empleados en la consulta y los documentos donde aparecen:

tf_{td} (número de veces que aparece un término en un documento)[casilla vacía implica un cero]

	DIGITALIZACIÓN	DOCUMENTOS	BIBLIOTECAS
D1	1	1	1
D2			1
D3	1		1

Ahora calculamos los valores de $TF_{td} = 1 + \log(1 + \log(tf_{td}))$ [casilla vacía implica un cero]

	DIGITALIZACIÓN	DOCUMENTOS	BIBLIOTECAS
D1	1	1	1
D2			1
D3	1		1

Los valores de IDF de cada uno de los términos empleados en la consulta son los siguientes:

IDF (digitalización) = $\log \frac{4+1}{2}$ = 0’40

IDF (documentos) = $\log \frac{4+1}{1}$ = 0’70

IDF (bibliotecas) = $\log \frac{4+1}{3}$ = 0’22

Los valores de los pesos de los términos en los documentos se resumen en la siguiente tabla (casilla vacía implica un cero):

W_{td}

	digitalización	documentos	bibliotecas
D1	0’4	0’7	0’22
D2			0’22
D3	0’4		0’22

Para calcular el peso de los términos en la consulta, consideramos exclusivamente su frecuencia de aparición:

W_{tq}

	digitalización	documentos	bibliotecas
Q	3	1	2

El cálculo de los parámetros de normalización 'K' correspondientes a cada documento en el que aparece alguno de los términos de la consulta da los siguientes valores:

K(D1) = $\frac{1}{(1-0,2)+0,2.\frac{167}{245}} = \frac{1}{0,94}$ = 1'06

K(D2) = $\frac{1}{(1-0,2)+0,2.\frac{238}{245}} = \frac{1}{0,99}$ = 1'01

K(D3) = $\frac{1}{(1-0,2)+0,2.\frac{306}{245}} = \frac{1}{1,05}$ = 0'95

Finalmente, el cálculo de similaridades entre los documentos de la colección y la consulta es el siguiente:

Sim(D1,Q) = K(D1). w(digitalización,D1).w(digitalización,Q) +

K(D1). w(documentos,D1).w(documentos,Q) + K(D1). (bibliotecas,D1).w(bibliotecas,Q) = 1'06 . 0'4 . 3 + 1'06 . 0'7 . 1 + 1'06 . 0'22 . 2 = 1'27 + 0'74 + 0'47 = 2'48

Sim(D2,Q) = K(D2). w(bibliotecas,D2).w(bibliotecas,Q) = 1'01 . 0'22 . 2 = 0'44

Sim(D3,Q) = K(D3). w(digitalización,D3).w(digitalización,Q) +

K(D3). w(bibliotecas,D3).w(bibliotecas,Q) = 0'95 . 0'4 . 3 + 0'95 . 0'22 . 2 = 1'14 + 0'42 =

= 1'56

El documento D4 de la colección tienen un valor de similaridad 0 con la consulta Q, dado que no tiene ningún término en común con la consulta, por lo que figurará en último lugar en la respuesta.

Por tanto, la respuesta del Sistema a Q sería la siguiente:

D1 / D3 / D2 / D4

EJERCICIO 4

Sea un Sistema de Recuperación de Información cuya matriz de ocurrencias término/documento es la siguiente (solamente se incluyen los 6 primeros documentos y una selección de 12 términos):

	D1	D2	D3	D4	D5	D6
difusión	0	1	1	0	3	1
inestabilidad	1	0	1	2	0	0
información	0	1	2	0	0	1
internet	0	0	0	1	1	0
obsolescencia	1	2	1	1	0	0
precisión	0	0	1	0	0	0
profundo	1	0	0	1	0	1
semántica	1	1	2	0	2	1
superficial	2	0	0	1	0	2
web	0	0	1	3	1	0
wide	1	1	0	2	0	0
world	1	0	1	2	0	0

La colección consta de 100000 documentos (N=100000).

El número ‘n’ de documentos de la colección en los que aparecen los términos de la consulta se puede resumir en la siguiente tabla:

	n
obsolescencia	1069
world	2308
wide	1847
web	3115

Considere el valor habitual para el parámetro ’s’:

s = 0’2

La longitud de los documentos de la colección es la siguiente:

D1 = 1761 bytes; D2 = 2830 bytes; D3 = 6613 bytes;

D4 = 345 bytes; D5 = 522 bytes; D6 = 4273 bytes;

Considere que la longitud media de los documentos de la colección es de 520 bytes:

avgdl = 520 bytes

Calcule el peso de los términos en la consulta como si se tratase de un documento más, empleando el IDF del término en la colección. Los números entre paréntesis señalan la frecuencia del término correspondiente en la consulta. Así, información(3) indicaría que el término ‘información’ aparece 3 veces en la consulta.

Es suficiente utilizar dos decimales en los cálculos. A la hora de representar la respuesta, se indicará con el símbolo '/' el orden entre documentos, y una coma servirá para indicar que el sistema no puede imponer un orden entre documentos (debido a que poseen la misma similaridad con la consulta). Así, la respuesta D3 / D5 / D2, D4 / D1 indicaría que el sistema situaría en primer lugar, con la máxima similaridad, el documento D3; a continuación, con menor similaridad, el documento D5; a continuación, con la misma similaridad –aunque menor que la que posee el documento D5-, el sistema incluye los documentos D2 y D4; por último, con la menor similaridad, el documento D1.

Hallar la respuesta inicial del sistema en relación a la consulta:

Q = obsolescencia(2) world(1) wide(1) web(1)

SOLUCIÓN

Dado que se nos da en el enunciado la matriz de ocurrencias término/documento, podemos calcular los valores de TF_{td} . En la siguiente tabla incluimos solamente los términos que aparecen en la consulta:

TF_{td} = 1 + log(1 + log(tf_{td})) [casilla vacía implica un cero]

	obsolescencia	world	wide	web
D1	1	1	1	
D2	1’11		1	
D3	1	1		1
D4	1	1’11	1’11	1’17
D5				1
D6				

Los valores de IDF de cada uno de los términos empleados en la consulta son los siguientes:

IDF (obsolescencia) = log $\frac{100000+1}{1069}$ = log 93’55 = 1’97

IDF (world) = log $\frac{100000+1}{2308}$ = log 43’33 = 1’64

IDF (wide) = log $\frac{100000+1}{1847}$ = log 54’14 = 1’73

IDF (web) = log $\frac{100000+1}{3115}$ = log 32’10 = 1’51

Los valores de los pesos de los términos en los documentos se resumen en la siguiente tabla (casilla vacía implica un cero):

W_{td}

	OBSOLESCENCIA	WORLD	WIDE	WEB
D1	1’97	1’64	1’73	
D2	2’19		1’73	
D3	1’97	1’64		1’51
D4	1’97	1’82	1’92	1’77
D5				1’51
D6				

Para calcular el peso de los términos en la consulta, consideramos inicialmente su frecuencia de aparición:

tf_{tq} (número de veces que aparece un término en un documento)

	OBSOLESCENCIA	WORLD	WIDE	WEB
Q	2	1	1	1

Al considerar la consulta como un documento más, calculamos a continuación los valores de $TF_{tq} = 1 + \log(1 + \log(tf_{tq}))$

	OBSOLESCENCIA	WORLD	WIDE	WEB
Q	1’11	1	1	1

Con estos valores, calculamos los pesos de los términos en la consulta empleando los IDF de los términos en la colección:

$W_{tq} = TF_{tq} \cdot IDF_t$

	OBSOLESCENCIA	WORLD	WIDE	WEB
Q	2’19	1’64	1’73	1’51

El cálculo de los parámetros de normalización ‘K’ correspondientes a cada documento en el que aparece alguno de los términos de la consulta da los siguientes valores:

K(D1) = $\frac{1}{(1-0,2)+0,2.\frac{1761}{520}} = \frac{1}{1,48}$ = 0’68

K(D2) = $\frac{1}{(1-0,2)+0,2.\frac{2830}{520}} = \frac{1}{1,89}$ = 0’53

K(D3) = $\frac{1}{(1-0,2)+0,2.\frac{6613}{520}} = \frac{1}{3,34}$ = 0’30

K(D4) = $\frac{1}{(1-0,2)+0,2.\frac{345}{520}} = \frac{1}{0,93}$ = 1’08

K(D5) = $\frac{1}{(1-0,2)+0,2.\frac{522}{520}} = \frac{1}{1}$ = 1

K(D6) = $\frac{1}{(1-0,2)+0,2.\frac{4273}{520}} = \frac{1}{2,44}$ = 0’41

Finalmente, el cálculo de similaridades entre los documentos de la colección y la consulta es el siguiente:

Sim(D1,Q) = K(D1). w(obsolescencia,D1).w(obsolescencia,Q) + K(D1). w(world,D1).w(world,Q)

+ K(D1). w(wide,D1).w(wide,Q) = 0’68 . 1’97 . 2’19 + 0’68 . 1’64 . 1’64 +

+ 0’68 . 1’73 . 1’73 = 2’93 + 1’83 + 2’04 = 6’80

Sim(D2,Q) = K(D2). w(obsolescencia,D2).w(obsolescencia,Q) + K(D2). w(wide,D2).w(wide,Q) =

= 0’53 . 2’19 . 2’19 + 0’53 . 1’73 . 1’73 = 2’54 + 1,59 = 4’13

Sim(D3,Q) = K(D3). w(obsolescencia,D3).w(obsolescencia,Q) + K(D3). w(world,D3).w(world,Q)

+ K(D3). w(web,D3).w(web,Q)=

= 0’30 . 1’97 . 2’19 + 0’30 . 1’64 . 1’64 + 0’30 . 1’51 . 1’51 = 1’29 + 0’81 + 0’68 =

= 2’78

Sim(D4,Q) = K(D4). w(obsolescencia,D4).w(obsolescencia,Q) + K(D4). w(world,D4).w(world,Q)

+ K(D4). w(wide,D4).w(wide,Q) + K(D4). w(web,D4).w(web,Q) =

= 1’08 . 1’97 . 2’19 + 1’08 . 1’82 . 1’64 + 1’08 . 1’92 . 1’73 + 1’08 . 1’77 . 1’51 =

= 4’66 + 3’22 + 3’59 + 2’89 = 14’36

Sim(D5,Q) = K(D5). w(web,D5).w(web,Q) = 1 . 1’51 . 1’51 = 2’28

Sim(D6,Q) = 0

En relación a los seis primeros documentos de la colección, la respuesta del Sistema a Q sería la siguiente:

D4 / D1 / D2 / D3 / D5 / D6

EJERCICIO 5

Sea un Sistema de Recuperación de Información cuya colección consta de 3 millones de documentos. La información relativa a los primeros cuatro documentos de dicha colección (los términos de indexación que contienen, la frecuencia de aparición y sus posiciones) se resume en la siguiente tabla:

	TÉRMINOS(FRECUENCIA,<POSICIONES>)
D1	W3C(1,<61>), normalización(1,<46>), internet(2,<16,503>), 1994(1,<4>), Berners-Lee(1,<7>), objetivos(1,<73>), semántica (1,<12>), accesible(1,<91>)
D2	W3C(1,<52), recomendaciones(1,<13>), estándares(2,<52,66>), interoperabilidad(1,<93>), tecnologías(2,<31,40>), web(2,<42,88>), internet(1,<125>), ontologías(1,<55>)
D3	W3C(2,<6,93>), recomendaciones(2,<41,91>), RDF(2,<109,218>), ontologías(1,<47>), especificaciones(1,<2>), lenguaje(1,<44>)
D4	RDF(1,<21>), RDFS(1,<23>), XML(1,<80>), sintaxis(2,<79>), lenguaje(2,<113,160>), ontologías(2,<116,163>), concepto(1,<40>)

En la consulta se han empleado los siguientes términos: normalización, lenguaje y ontologías.

El número 'n' de documentos de la colección en los que aparecen los términos de la consulta se puede resumir en la siguiente tabla:

	n
normalización	1634
lenguaje	898
ontologías	1425

El usuario ha impuesto pesos a los términos de la consulta, que una vez normalizados son los siguientes (entre paréntesis):

$$Q = \text{normalización}(3'21) \quad \text{lenguaje}(1'07) \quad \text{ontologías}(2'94)$$

Considere el valor habitual para el parámetro 's':

$$s = 0'2$$

La longitud de los documentos de la colección es la siguiente:

$$D1 = 1543 \text{ bytes}; D2 = 1671 \text{ bytes}; D3 = 2381 \text{ bytes}; D4 = 694 \text{ bytes}$$

Considere que la longitud media de los documentos de la colección es de 464 bytes:

$$avgdl = 464 \text{ bytes}$$

Es suficiente utilizar dos decimales en los cálculos. A la hora de representar la respuesta, se indicará con el símbolo '/' el orden entre documentos, y una coma servirá para indicar que el sistema no puede imponer un orden entre documentos (debido a que poseen la misma similaridad con la consulta). Así, la respuesta D3 / D5 / D2, D4 / D1 indicaría que el sistema

situaría en primer lugar, con la máxima similaridad, el documento D3; a continuación, con menor similaridad, el documento D5; a continuación, con la misma similaridad –aunque menor que la que posee el documento D5-, el sistema incluye los documentos D2 y D4; por último, con la menor similaridad, el documento D1.

El ejercicio consiste en hallar la respuesta del sistema en relación a los primeros cuatro documentos de la colección y ante la consulta anteriormente señalada:

Q = normalización(3’21) lenguaje(1’07) ontologías(2’94)

SOLUCIÓN

Resumimos la matriz de ocurrencias término/documento en la siguiente tabla, incluyendo solamente los términos empleados en la consulta y los documentos donde aparecen:

tf_{td} (número de veces que aparece un término en un documento)[casilla vacía implica un cero]

	NORMALIZACIÓN	LENGUAJE	ONTOLOGÍAS
D1	1		
D2			1
D3		1	1
D4		2	2

Ahora calculamos los valores de $TF_{td} = 1 + \log(1 + \log(tf_{td}))$ [casilla vacía implica un cero]

	NORMALIZACIÓN	LENGUAJE	ONTOLOGÍAS
D1	1		
D2			1
D3		1	1
D4		1’11	1’11

Los valores de IDF de cada uno de los términos empleados en la consulta son los siguientes:

IDF (normalización) = $\log \frac{3000000+1}{1634}$ = log 1835’99 = 3’26

IDF (lenguaje) = $\log \frac{3000000+1}{898}$ = log 3340’76 = 3’52

IDF (ontologías) = $\log \frac{3000000+1}{1425}$ = log 2105’26 = 3’32

Los valores de los pesos de los términos en los documentos se resumen en la siguiente tabla (casilla vacía implica un cero):

W_{td}

	normalización	lenguaje	ontologías
D1	3’26		
D2			3’32
D3		3’52	3’32
D4		3’91	3’69

No es preciso calcular el peso de los términos en la consulta, pues vienen impuestos por el usuario. Basta considerar los valores incluidos en el enunciado:

W_{tq}

	normalización	lenguaje	ontologías
Q	3’21	1’07	2’94

El cálculo de los parámetros de normalización ‘K’ correspondientes a cada documento en el que aparece alguno de los términos de la consulta da los siguientes valores:

K(D1) = $\frac{1}{(1-0,2)+0,2.\frac{1543}{464}} = \frac{1}{1,47}$ = 0’68

K(D2) = $\frac{1}{(1-0,2)+0,2.\frac{1671}{464}} = \frac{1}{1,52}$ = 0’66

K(D3) = $\frac{1}{(1-0,2)+0,2.\frac{2381}{464}} = \frac{1}{1,83}$ = 0’55

K(D4) = $\frac{1}{(1-0,2)+0,2.\frac{694}{464}} = \frac{1}{1,10}$ = 0’91

Finalmente, el cálculo de similaridades entre los documentos de la colección y la consulta es el siguiente:

Sim(D1,Q) = K(D1). w(normalización,D1).w(normalización,Q) = 0’68 . 3’26 . 3’21 = 7’12

Sim(D2,Q) = K(D2). w(ontologías,D2).w(ontologías,Q) = 0’66 . 3’32 . 2’94 = 6’44

Sim(D3,Q) = K(D3). w(lenguaje,D3).w(lenguaje,Q) + K(D3). w(ontologías,D3).w(ontologías,Q) =

= 0’55 . 3’52 . 1’07 + 0’55 . 3’32 . 2’94 = 2’07 + 5’37 = 7’44

Sim(D4,Q) = K(D4). w(lenguaje,D4).w(lenguaje,Q) + K(D4). w(ontologías,D4).w(ontologías,Q) =

= 0’91 . 3’91 . 1’07 + 0’91 . 3’69 . 2’94 = 3’81 + 9’87 = 13’68

Por tanto, la respuesta del Sistema en relación a los primeros cuatro documentos de la colección y ante la consulta Q sería la siguiente:

D4 / D3 / D1 / D2

EJERCICIO 6

Sea un Sistema de Recuperación de Información cuya colección consta de 6 millones de documentos. La colección incluye 1037489 términos de indexación, de los cuales los términos incluidos en las últimas consultas de un usuario tienen la siguiente distribución:

	N
T49	741509
T102	1563802
T2431	1299347
T32860	577693
T572319	238414

Siendo 'n' el número de documentos de la colección en los que aparece el término.

Cinco de los documentos de la colección están representados de la siguiente manera, en relación a los términos indicados anteriormente:

	T49	T102	T2431	T32860	T572319
D46	0	1	0	1	0
D703	0	0	0	1	0
D5143	2	0	1	0	0
D38091	1	0	2	0	1
D610102	0	1	0	0	2

El usuario ha impuesto pesos a los términos de la consulta, que una vez normalizados son los siguientes (entre paréntesis):

Q = T2431(1'82) T32860(2'05) T49(1'43)

Considere el valor habitual para el parámetro 's':

s = 0'2

La longitud de los documentos de la colección es la siguiente:

D46 = 43 bytes; D703 = 71 bytes; D5143 = 81 bytes; D38091 = 94 bytes; D610102 = 47 bytes

Considere que la longitud media de los documentos de la colección es de 75 bytes:

avgdl = 75 bytes

Es suficiente utilizar dos decimales en los cálculos. A la hora de representar la respuesta, se indicará con el símbolo '/' el orden entre documentos, y una coma servirá para indicar que el sistema no puede imponer un orden entre documentos (debido a que poseen la misma similaridad con la consulta). Así, la respuesta D3 / D5 / D2, D4 / D1 indicaría que el sistema situaría en primer lugar, con la máxima similaridad, el documento D3; a continuación, con menor similaridad, el documento D5; a continuación, con la misma similaridad –aunque menor

que la que posee el documento D5-, el sistema incluye los documentos D2 y D4; por último, con la menor similaridad, el documento D1.

Hallar la respuesta del sistema en relación a los documentos de la colección seleccionados y ante la consulta anteriormente señalada:

Q = T2431(1’82) T32860(2’05) T49(1’43)

SOLUCIÓN

Dado que se nos da en el enunciado la matriz de ocurrencias término/documento, podemos calcular los valores de TF_{td} . En la siguiente tabla incluimos solamente los términos que aparecen en la consulta:

$TF_{td} = 1 + \log(1 + \log(tf_{td}))$ [casilla vacía implica un cero]

	T2431	T32860	T49
D46		1	
D703		1	
D5143	1		1’11
D38091	1’11		1
D610102			

Los valores de IDF de cada uno de los términos empleados en la consulta son los siguientes:

IDF (T2431) = $\log \frac{6000000+1}{1299347}$ = log 4’62 = 0’66

IDF (T32860) = $\log \frac{6000000+1}{577693}$ = log 10’39 = 1’02

IDF (T49) = $\log \frac{6000000+1}{741509}$ = log 8’09 = 0’91

Los valores de los pesos de los términos en los documentos se resumen en la siguiente tabla (casilla vacía implica un cero):

W_{td}

	T2431	T32860	T49
D46		1’02	
D703		1’02	
D5143	0’66		1’01
D38091	0’73		0’91
D610102			

No es preciso calcular el peso de los términos en la consulta, pues vienen impuestos por el usuario. Basta considerar los valores incluidos en el enunciado:

W_{tq}

	T2431	T32860	T49
Q	1’82	2’05	1’43

El cálculo de los parámetros de normalización 'K' correspondientes a cada documento en el que aparece alguno de los términos de la consulta da los siguientes valores:

K(D46) = $\frac{1}{(1-0,2)+0,2.\frac{43}{75}} = \frac{1}{0,91}$ = 1'10

K(D703) = $\frac{1}{(1-0,2)+0,2.\frac{71}{75}} = \frac{1}{0,99}$ = 1'01

K(D5143) = $\frac{1}{(1-0,2)+0,2.\frac{81}{75}} = \frac{1}{1,02}$ = 0'98

K(D38091) = $\frac{1}{(1-0,2)+0,2.\frac{94}{75}} = \frac{1}{1,05}$ = 0'95

Finalmente, el cálculo de similaridades entre los documentos de la colección y la consulta es el siguiente:

Sim(D46,Q) = K(D46). w(T32860,D46).w(T32860,Q) = 1'1 . 1'02 . 2'05 = 2'30

Sim(D703,Q) = K(D703). w(T32860,D703).w(T32860,Q) = 1'01 . 1'02 . 2'05 = 2'11

Sim(D5143,Q) = K(D5143). w(T2431,D5143).w(T2431,Q) + K(D5143). w(T49,D5143).w(T49,Q) =

= 0'98 . 0'66 . 1'82 + 0'98 . 1'01 . 1'43 = 1'18 + 1'42 = 2'60

Sim(D38091,Q) = K(D38091). w(T2431,D38091).w(T2431,Q)+

+ K(D38091). w(T49,D38091).w(T49,Q) = 0'95 . 0'73 . 1'82 + 0'95 . 0'91 . 1'43 =

= 1'26 + 1'24 = 2'50

Sim(D610102,Q) = 0 [no existe ningún término de la consulta en este documento]

Por tanto, la respuesta del Sistema en relación a los cinco documentos de la colección seleccionados y ante la consulta Q sería la siguiente:

D5143 / D38091 / D46 / D703 / D610102