

**Sistemas Inteligentes de Acceso a la Información**  
**Problemas y cuestiones**  
**Problemas y mejoras en la Recuperación**  
**de Información sobre texto**

1. Suponiendo que en respuesta a una consulta “dog race”, y usando realimentación por relevancia (algoritmos de Rocchio, Ide\_Regular e Ide\_Dec\_Hi), un usuario marca como relevante el primer documento “greyhound race track betting” y como no relevantes los documentos segundo y tercero, “Iditarod dog sled race” y “Husky dog sled race”, calcular el vector de pesos de términos para la consulta realimentada. Asumir que la representación de los documentos es binaria y que  $\alpha = \beta = \gamma = 2$ . Ordenar los términos alfabéticamente.
2. Sea la consulta “a b” (con pesos binarios) y supongamos que se recuperan los documentos 1 a 6 de una colección determinada. Si la representación de los documentos es la siguiente:

D	a	b	c	d	e
1	1	0	0	1	1
2	1	0	1	1	0
3	1	0	1	0	0
4	0	1	1	1	1
5	0	1	1	1	1
6	0	1	0	0	0

Si se aplica una expansión de consulta por agrupamiento sobre los documentos relevantes, utilizando la fórmula del coseno para el cálculo de similitud, definir cual es la nueva consulta si el criterio de inclusión de un término “t” en el grupo de un término “x” de la consulta es:

- a.  $SIM(t,x) > 0.5$
  - b.  $SIM(t,x) > 0.1$
  - c. t es el término más similar a x
3. Supongamos que se dispone de la colección de documentos del problema anterior, y se desea crear un thesaurus estadístico utilizando el método HAC y la similitud del coseno, con el método del centroide. Representar el dendograma obtenido y cuales son las clases obtenidas si se fija el número de grupos  $|L| = 2$  y a 3.
  4. Sea la siguiente matriz de similitudes entre términos para el vocabulario de una colección dada:

	a	b	c	d	e
a	1	0,5	0,2	0	0
b	0,5	1	0,1	0,2	0
c	0,2	0,1	1	0	0
d	0	0,2	0	1	0,7
e	0	0	0	0,7	1

Se desea crear un thesaurus estadístico utilizando el método HAC y la similitud del coseno. Representar el dendograma obtenido y cuales son las clases obtenidas si se fija el número de grupos  $|L| = 2$  y a 3, si se utiliza:

- a. El método de enlace simple.
- b. El método de enlace completo.

¿Es posible aplicar el método del centroide y el de media de grupo con la información disponible?

5. Sea el problema consistente en clasificar un documento como Biología o Química, y sean los siguientes documentos de entrenamiento:

Biología: “ADN célula”

Biología: “ADN mitocondria”

Química: “hidrógeno molécula”

Química: “oxígeno átomo”

Calcular la clasificación del documento “ADN molécula” usando un clasificador de Rocchio sobre una representación con pesos  $tf$  y con  $\beta = \gamma = 1$ .

6. Sean los documentos del problema 2. Se desea crear un agrupamiento en dos temas usando el algoritmo K-Means, usando como fórmula de similitud el producto escalar de vectores, y realizando 3 iteraciones. Si los puntos de inicio son los propios documentos D1 y D4, ¿cuáles son los grupos finales? ¿Y si los puntos de inicio son D1 y D3?