

**Sistemas Inteligentes de Acceso a la Información**  
**Problemas y cuestiones**  
**Técnicas básicas de Recuperación**  
**de Información sobre texto**

1. Se asume que el número total de documentos de una colección es de 10,000, y que las siguientes palabras aparecen en los siguientes números de documentos:

“cell” aparece en 500 documentos

“DNA” aparece en 100 documentos

“molecule” aparece en 200 documentos

“one” aparece en 10,000 documentos

Sabiendo que en la representación de los documentos se utilizan esos cuatro términos, y pesos tipo TF.IDF, se pide calcular el vector de pesos de términos para el documento: “cell molecule DNA one cell DNA”.

2. Asumiendo pesos tipo TF en la representación de los siguientes documentos, calcular la similitud entre ellos por medio de la fórmula del coseno:

“hotel Maui discount hotel beach”

“Maui hotel Maui luxury hotel five star”

“hotel”

Calcular también la similitud por medio del producto escalar simple, y de la inversa de la distancia euclídea. Comparar las tres similitudes y discutir la conveniencia de usar una u otra en un entorno práctico.

3. Ilustrar gráficamente con una tabla simple de listas enlazadas el índice inverso construido para la colección formada por los tres siguientes documentos, asumiendo indexación TF. Ordenar las palabras alfabéticamente.

Doc1: “dog pet dog food dog collar”

Doc2: “pet cat food cat litter pet”

Doc3: “fish food turtle food”

4. Para los documentos del problema anterior, y la consulta “pet food”, calcular las similitudes entre la consulta y los documentos, y el ranking resultante. Se asume representación con pesos TF y se utiliza la similitud del coseno.
5. Sea la consulta “(pet  $\wedge$  food)  $\vee$   $\neg$ fish”. Construir la fórmula DNF asociada a la consulta. Representar como fórmulas los documentos del problema 3, y aplicar el modelo booleano para calcular la similitud entre cada documento y la consulta. ¿Es posible construir un ranking? Desarrollar posibilidades para este aspecto.
6. Sea el documento  $d = \text{“gcatcgagagagtatacagtagc”}$  y la secuencia  $w = \text{“gcagagag”}$ , en un problema de reconocimiento de ADN Se pide calcular el número de intentos de ajuste y de comparaciones de caracteres realizadas por el algoritmo de búsqueda ingenua y el algoritmo de Boyer-Moore Simplificado. Comparar ambos números y discutir la bondad de los algoritmos.