

5) (**) Se tienen los siguientes documentos:

D1: CORDERO SAL PIMIENTA ROMERO

D2: CERDO CORDERO SAL CORDERO

D3: SAL CERDO LIMON

D4: CORDERO ENTRAÑA

D5: PIMIENTA PAPA CORDERO PIMIENTA

D6: CORDERO CORDERO CORDERO CORDERO

Dada la consulta "CORDERO PIMIENTA" dar el resultado de la consulta rankeada utilizando TF.IDF. (10 pts)

Considerando como relevante los documentos que no tengan otra carne que no sea CORDERO, calcular la Precisión, Recall y F1 Score. (5 pts)

Primero calculamos el IDF de cada término de la consulta en particular: $IDF = \log(N+1/ft_i)$

- cordero: $IDF = \log(7/5) = 0.14$
- pimienta: $IDF = \log(7/2) = 0.54$

No tomamos en cuenta los documentos en los que no aparece ninguno de los dos términos. Teniendo que el TF es la cantidad de veces que aparece el término en determinado documento, el cálculo del ranking queda de la forma:

D1: $1 * 0.14 + 1 * 0.54 = 0.68$

D2: $2 * 0.14 = 0.28$

D4: $1 * 0.14 = 0.14$

D5: $1 * 0.14 + 2 * 0.54 = 1.22$

D6: $4 * 0.14 = 0.56$

Entonces el ranking queda: D5, D1, D6, D2, D4.

Los documentos que consideramos relevantes son: D1, D5, D6

| | Doc recuperados | No recuperados |
|---------------|-----------------|----------------|
| Relevantes | a | b |
| No relevantes | c | d |

Recordando:

precisión = $a / a + c$

recall = $a / a + b$

F1 score = $2 \text{ precisión} * \text{recall} / \text{precision} + \text{recall}$

Usando estas fórmulas con la consulta anterior obtenemos los valores de a, b, c y d y los reemplazamos en dichas fórmulas.

$$\text{precisión} = 3 / 5$$

$$\text{recall} = 3 / 3 = 1$$

$$\text{F1 score} = 2 * (3 / 5) * 1 / (3 / 5 + 1) = 6 / 5 / 8/5 = \mathbf{3/4}$$