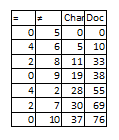
**CONSIGNA**

4) A partir de la siguiente estructura de un índice Invertido:



madreigueragistradomaleficiotanantialmanifiesto

0->1010010011

10->01101010010010101011111

33->01011

38->11101010110101010

55->01001001000100

69->0111010

76->0101011010100100

(las posiciones en los documentos están numeradas desde la posición 1, y todo está codificado en gamma)

Armado con la finalidad de poder resolver consultas por proximidad, se pide extraer la información de los documentos y resolver la consulta “**Madriguera Maleficio** ” utilizando **TF-IDF**. Detallar cada paso realizado (\*\*\*\*) (20pts)

**RESOLUCIÓN**

Se sabe por la tabla utilizada que el formato utilizado para almacenar los términos es con front coding parcial. Asumiendo esto, entonces se sabe que los términos son:

Madre, Madriguera, Magistrado, Maleficio, Maleta, Manantial, Manifiesto

Si deciframos además los códigos gamma, con estructura para encontrar consultar por proximidad (Formato: Termino Documento Frecuencia Posiciones), obtenemos:

* Madre: 1 010 010 011 -> 1, 2, 2, 3 -> Doc 1, Freq 2, Pos 2, Pos 3 -> D1:2(2,3)
* Madriguera: 011 010 1 00100 1 010 1 011 1 1 1 ->

-> Doc 3, Freq 2, Pos 1, Pos 4; Doc 4, Freq 2, Pos 1, Pos 3; Doc 5, Freq 1, Pos 1

D3:2(1,4) ; D4:2(1,3) ; D5:1(1)

* Magistrado: 010 1 1 -> D2:1(1)
* Maleficio: 1 1 1 010 1 011 010 1 010

D1:1(1) ; D3:1(3) ; D5:1(2)

* Maleta: 010 010 010 00100

D2:2(2,4)

* Manantial: 011 1 010

D3:1(2)

* Manifiesto: 010 1 011 010 1 00100

D2:1(3) ; D4:1(4)

Ahora entonces conozco como se distribuyen los términos en cada documento. Entonces si se hace el Query de “Madriguera Maleficio” tenemos que poseer una tabla con los TF y los IDF de cada término:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | TF | | | | | IDF |
| D1 | D2 | D3 | D4 | D5 |
| Madre | 2 | - | - | - | - | 0,7782 |
| Madriguera | - | - | 2 | 2 | 1 | 0,3010 |
| Magistrado | - | 1 | - | - | - | 0,7782 |
| Maleficio | 1 | - | 1 | - | 1 | 0,3010 |
| Maleta | - | 2 | - | - | - | 0,7782 |
| Manantial | - | - | 1 | - | - | 0,7782 |
| Manifiesto | - | 1 | - | 1 | - | 0,4771 |

Entonces haciendo el Query obtenemos:

* D1: 1\*0,3010 = 0,3010
* D2: 0 = 0
* D3: 2\*0,3010+1\*0,3010 = 0,9030
* D4: 2\*0,3010 = 0,6020
* D5: 1\*0,3010+1\*0,3010 = 0,6020

El Rank de consultas finalmente quedaría:

|  |
| --- |
| **RANKING “MADRIGUERA MALEFICIO”** |
| D3 = 0,9030 |
| D4 = 0,6020 |
| D5 = 0,6020 |
| D1 = 0,3010 |
| D2 = 0 |