**CONSIGNA 1ER PARCIAL 2019 2C**

5) Dados los documentos:

D1: AUTO AUTO AUTO CAMIONETA CAMION MOTO MOTO

D2: CAMION MOTO

D3: AUTO CAMIONETA

D4: AUTO SUV CAMION CAMIONETA MOTO

D5: SUV

a) Resolver la consulta ranqueada “AUTO SUV” utilizando TF-IDF. Detallar cada paso para llegar al resultado (\*\*\* 7 pts)

b) Proponer una alternativa de solución que mejore la obtenida en el pto a. Justificar la respuesta.. (\*\*\* 4 pts)

c) Considerando relevantes los documentos que contienen los términos AUTO y SUV, calcular la precisión, recall, y F1 score para el resultado obtenido en el punto a. (\*\* 4 pts)

**RESOLUCIÓN**

1. Para resolver la consulta Auto Suv, primero tengo que listar los términos posibles, y los documentos en los cuales aparecen, y con que frecuencia:

* Auto: D1:3(1,2,3), D3:1(1), D4:1(1)
* Camión: D1:1(5), D2:1(1), D4:1(3)
* Camioneta: D1:1(4), D3:1(2), D4:1(4)
* Moto: D1:2(6,7), D2:1(2), D4:1(5)
* Suv: D4:1(2), D5:1(1)

Entonces ahora, puedo armar la tabla de TF-IDF requerida para la consulta ranqueada:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | TF | | | | | IDF |
| D1 | D2 | D3 | D4 | D5 |
| Auto | 3 | 0 | 1 | 1 | 0 | 0,3010 |
| Camión | 1 | 1 | 0 | 1 | 0 | 0,3010 |
| Camioneta | 1 | 0 | 1 | 1 | 0 | 0,3010 |
| Moto | 2 | 1 | 0 | 1 | 0 | 0,3010 |
| Suv | 0 | 0 | 0 | 1 | 1 | 0,4771 |

Teniendo la tabla puedo hacer la consulta de cuales poseen los documentos y ranquearlos:

* D1: 3\*0,3010 = 0,9030
* D2: 0
* D3: 1\*0,3010 = 0,3010
* D4: 1\*0,3010 + 1\*0,4771 = 0,7781
* D5: 1\*0,4771 = 0,4771

Finalmente el ranking queda:

|  |
| --- |
| **RANKING DE “AUTO SUV”** |
| D1 = 0,9030 |
| D4 = 0,7781 |
| D5 = 0,4771 |
| D3 = 0,3010 |
| D2 = 0 |

1. Si se busca mejorar la solución del punto a se puede utilizar el método de BM25, que toma en consideración la longitud de los documentos para calcular la frecuencia del término en el mismo.

De esta forma, se la da aún más importancia a la aparición del término en el documento, basándose en la longitud del mismo. Si el término crece mucho además, el TF de BM25 crece de forma logarítmica por lo que la curva se aplanaría y si un término esta en excesiva cantidad de veces, no rankearía a un documento por encima de los demás solo por eso.

La fórmula que aplica al TF de BM25 es:

b = factor de normalización

|Dj| = Long del documento

Avg Dj = Longitud promedio de los documentos

Definiendo k podemos decir que buscamos que el TF indique.

* Si k -> infinito, entonces el TF tiende a ser el mismo que en TF-IDF
* Si k -> 0, entonces el TF es 1
* Dos valores de k muy utilizados son k=2 y k=1,2

Finalmente, lo que resta por definir es la fórmula final para BM25, que será:

Es decir:



Y considerando como relevantes únicamente los términos que contienen Auto y Suv (Únicamente D4 contiene a ambos), asumiendo Beta=1:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Precisión | Recall | F1 |
| D1 | 0 | 0 | - |
| D2 | 0 | 0 | - |
| D3 | 0 | 0 | - |
| D4(Relevante) | 1/4 | 1 | 0,4 |
| D5 | 1/5 | 1 | 0,3333 |