**Ranking Retrieval**

Profesor Heider Sanchez

**P1. Normalización de la longitud:**

Se tiene la siguiente tabla de conteos para la consulta y dos documentos.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | **Frequency ()** | | |
|  | **term** | **Q** | **Doc1** | **Doc2** |
|  | afecto | 115 | 15 | 30 |
|  | celosa | 10 | 5 | 0 |
|  | chisme | 2 | 20 | 22 |
|  | borrascoso | 0 | 25 | 0 |

Se le pide aplicar las siguientes técnicas de scoring:

1. El score es la sumatoria del *log-frecuency weigth* de los términos comunes.

En donde, y

1. El score ahora se aplica con los vectores unitarios.

En donde, , de igual forma para.

**Llene el siguiente cuadro,** analice los resultados y de una explicación de dicho comportamiento.

|  |  |  |
| --- | --- | --- |
|  | **(Q, Doc1)** | **(Q, Doc2)** |
| **DotProduct** |  |  |
| **Cosine** |  |  |
| **Explicación** |  | |

**P2. TF-IDF:**

Dada la siguiente tabla en donde se distribuye los pesos TF-IDF para dos documentos de la colección y para la consulta, se pide calcular el score sin normalizar (dot-product) y el score normalizado (cosine similarity) entre Q y cada documento.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Id | Término | Doc1 (TF-IDF) | Doc2 (TF-IDF) | Q (TF-IDF) |
| T1 | Clima | 1.452 | 0 | 0 |
| T2 | Biblioteca | 0 | 2.093 | 1.345 |
| T3 | Universidad | 2.122 | 0 | 1.453 |
| T4 | Alcalá | 3.564 | 0 | 1.987 |
| T5 | España | 4.123 | 4.245 | 0 |
| T6 | Libros | 0 | 1.234 | 2.133 |
| T7 | Geografía | 0 | 0 | 0 |
| T8 | Población | 2.342 | 0 | 0 |
| T9 | Electricidad | 0 | 0 | 0 |
| T10 | Ciencia | 0 | 0 | 0 |
| T11 | Social | 0 | 2.345 | 0 |
| T12 | Luz | 1.975 | 0 | 0 |
| T13 | Unamuno | 4.543 | 2.135 | 3.452 |
| T14 | Física | 0 | 0 | 0 |
| T15 | Fluidos | 6.134 | 0 | 0 |
| T16 | Literatura | 2.234 | 3.456 | 4.234 |

En donde *qi* es el peso tf-idf del término Ti respecto al documento Q. Lo mismo con di.

**Llene el siguiente cuadro,** analice los resultados y de una explicación de dicho comportamiento.

|  |  |  |
| --- | --- | --- |
|  | **(Q, Doc1)** | **(Q, Doc2)** |
| **DotProduct** |  |  |
| **Cosine** |  |  |