Repositorios de información

Recuperación de información / Seminarios

Daniel Gayo Avello

Ejercicio 1

Dada la siguiente **colección de documentos**:

```
d1 = “La pizza caprichosa es una pizza preparada con tomates,
     mozzarella, champiñones, aceitunas y corazones de alcachofa.”

d2 = “La pizza marinera es una pizza preparada con tomate, ajo,
     orégano y aceite de oliva virgen extra.”

d3 = “La pizza cuatro quesos es una pizza preparada con tomates,
     y los quesos mozzarella, gorgonzola, stracchino y fontina.”

d4 = “La pizza romana es una pizza preparada con tomates,
     mozzarella, anchoas y orégano.”

d5 = “La pizza vienesa es una pizza preparada con tomates,
     salchicha alemana, mozzarella, orégano y aceite de oliva
     virgen extra.”
```

Normalizar los documentos (paso a minúsculas y estematización) y generar el **vocabulario** para esa colección.

Ejercicio 2

Dada la colección de documentos del Ejercicio 1, construir la **matriz documento-término** (no el índice invertido) para un **modelo booleano** (es decir, sin ponderación de términos).

Ejercicio 3

Dada la matriz obtenida en el Ejercicio 2, **resolver las siguientes consultas.**

¡Atención! No se persigue obtener una lista de resultados sino determinar qué operaciones se deberían realizar sobre la matriz para obtener esa lista.

```
q1 = mozzarella AND tomates

q2 = aceitunas OR champiñones

q3 = mozzarella AND tomates AND NOT champiñones
```

Ejercicio 4

Dada la colección del ejercicio 1, calcule el **peso tf** (para cada término de cada documento) y el **peso idf** (para cada término de la colección).

Generar un **índice/archivo invertido** para esa colección que almacene esos pesos.

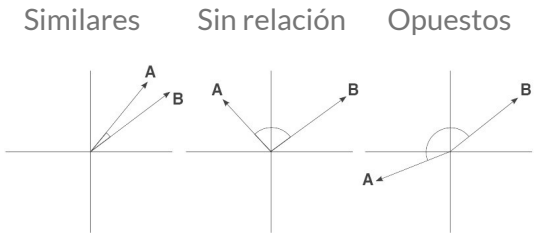
$$\mathrm{idf}(t, D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

Ejercicio 5

Dado el índice del Ejercicio 4, calcular la **similitud del coseno** para la **consulta** `"aceitunas tomate orégano"` con **cada uno de los documentos**.

A_i es el peso tf·idf de cada término de la consulta, B_i es el peso tf·idf de cada término del documento que se compara con la consulta.

$$\frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$



Ejercicio opcional: en lugar de la similitud del coseno, calcular la puntuación **Okapi BM25** para cada documento en relación con esa misma consulta.

Ejercicio 6

Supongamos una **colección de 100 documentos**: `d1..d100`

Para una consulta q el los **documentos relevantes** son:
`D* = {d2, d13, d43, d65, d89}`

Para esa consulta, un sistema de RI retorna los siguientes **resultados**:
`R = {d2, d13, d42, d65, d66, d88, d95, d43, d14, d89}`

Dada esa información, calcular la **precisión y exhaustividad** del sistema a diferentes niveles, p.ej., P@1, R@1, P@5, R@5, P@10, R@10.

¿Cómo se dibujaría una **curva de precisión vs exhaustividad**?