Información no estructurada

Práctica 3 -Modelos de IR

Javier Aróstegui Martín

Contenido

Implementación: 2
Modelo Vectorial: 2
RSJ (Modelo probabilístico): 2
BM25(Modelo probabilístico): 2
Rankings: 3

Implementación:

La implementación se ha realizado en Python con la ayuda de las librerías de pandas para el manejo de los datos y de numpy para operaciones matemáticas. Se han creado 3 ficheros: vectorial.py, bm25.py y rsj.py. Cada uno de ellos devuelve su ranking en función de las fórmulas que se presentan más adelante.

Modelo Vectorial:

Para el modelo vectorial he usado las siguientes fórmulas:

-Para el tf de los términos en los documentos:

$$tf(t,d) = \begin{cases} 1 + \log_2 frec(t,d) \text{ si } frec(t,d) > 0 \\ 0 \text{ en otro caso} \end{cases}$$

-Para el idf de los términos:

$$idf(t) = \log \frac{|\mathcal{D}| + 1}{|\mathcal{D}_t| + 0.5}$$

-Para el coseno:

$$\cos\left(\vec{d}, \vec{q}\right) = \frac{\vec{d} \cdot \vec{q}}{\left|\vec{d}\right|} = \frac{\sum_t d_t q_t}{\sqrt{\sum_t d_t^2}}$$

Esta fórmula no incluye la división entre el módulo de la consulta ya que, al ser común a todos los documentos, aunque se utilice no cambia el ranking. Es por esto por lo que en el ranking resultante los valores no se encuentran entre 0 y 1.

RSJ (Modelo probabilístico):

Para el cálculo de RSJ he utilizado la siguiente fórmula:

$$\text{RSJ}(d) = \sum_{w \in d \cap q} \log \frac{|\mathcal{D}| - |\mathcal{D}_w| + 0.5}{|\mathcal{D}_w| + 0.5}$$

Al disponer de muy pocos documentos he asignado |D| = 1000 como se sugiere en el enunciado para evitar valores negativos.

BM25(Modelo probabilístico):

En el caso de BM25 he utilizado la fórmula:

$$\sum_{w \in q} \frac{frec(w,d)}{|d|/\text{avg}_{d'}|d'| + frec(w,d)} \text{RSJ}(w)$$

Utilizando los cálculos de RSJ de la parte anterior.

Rankings:

Ranking Vectorial:

documento	
d15	1.337617
d20	1.182676
d3	1.051796
d14	0.912900
d11	0.888214
d7	0.796653
d8	0.779012
d19	0.574138
d13	0.547991
d6	0.493429
d12	0.474061
d2	0.474056
d4	0.444858
d5	0.428118
d1	0.335794
d18	0.318751
d9	0.313374
d10	0.274167
d16	0.268902
d17	0.244685

Ranking RSJ:

documento	
d8	25.425863
d18	25.425863
d2	25.425863
d4	25.425863
d15	25.425863
d7	25.425863
d13	25.425863
d12	25.425863
d11	25.425863
d20	21.273554
d1	21.133376
d19	20.879028
d16	20.879028
d3	20.879028
d17	20.879028
d9	20.777936
d6	20.777936
d5	20.777936
d10	20.777936
d14	16.625627

Ranking BM25:

documento	
d15	22.216659
d11	20.945189
d13	20.526941
d2	20.094230
d12	19.500320
d8	19.289066
d4	19.081992
d7	18.635036
d18	17.575258
d3	17.562696
d5	17.525806
d6	17.515666
d20	17.496136
d1	17.069818
d10	16.935225
d19	16.778567
d9	16.539637
d16	15.897233
d17	15.607245
d14	14.829460

El ranking RSJ no tiene en cuenta la frecuencia de los términos en los documentos y es por ello por lo que muchos documentos comparten valor (con que aparezca el mismo número de palabras de la consulta en el documento ya se obtiene el mismo valor). BM25 profundiza un poco más y a esos valores les aplica una operación que sí depende de la frecuencia y por tanto si creemos que el hecho de que aparezca más veces las palabras de la consulta hace que el documento sea más relevante, podemos asumir que BM25 es superior a RSJ. BM25 y el modelo vectorial posicionan al documento 15 el primero y aunque a partir del segundo ya no coinciden, los documentos más y menos relevantes de cada uno no se alejan demasiado. Por ejemplo, podemos ver como el documento 11 está posicionado bastante alto en ambos casos y los documentos 16 y 17 muy abajo.