Información no estructurada

Práctica 3 -Modelos de IR

Javier Aróstegui Martín

Contenido

[Implementación: 2](#_Toc67933737)

[Modelo Vectorial: 2](#_Toc67933738)

[RSJ (Modelo probabilístico): 2](#_Toc67933739)

[BM25(Modelo probabilístico): 2](#_Toc67933740)

[Rankings: 3](#_Toc67933741)

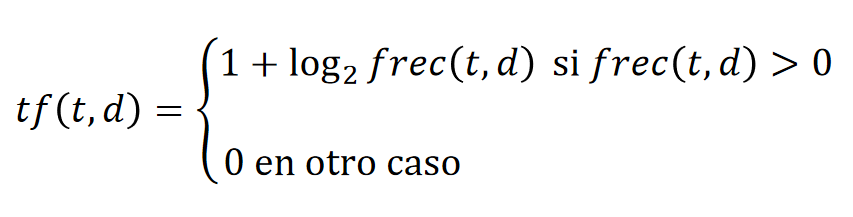
# Implementación:

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.

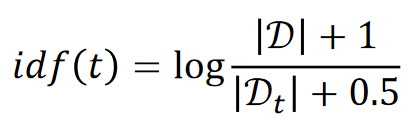
# Modelo Vectorial:

Para el modelo vectorial he usado las siguientes fórmulas:

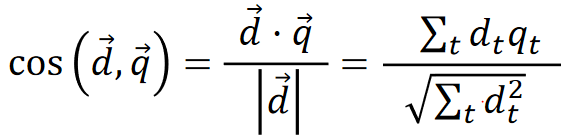
-Para el tf de los términos en los documentos:



-Para el idf de los términos:



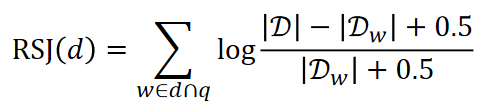
-Para el coseno:



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):

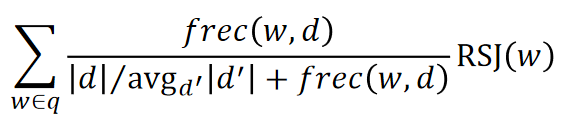
Para el cálculo de RSJ he utilizado la siguiente fórmula:



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

# BM25(Modelo probabilístico):

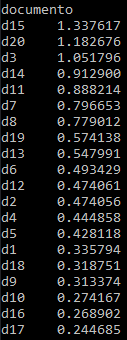
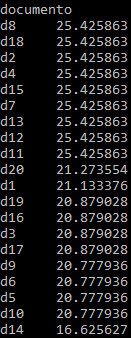
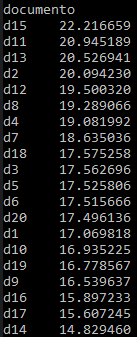
En el caso de BM25 he utilizado la fórmula:



Utilizando los cálculos de RSJ de la parte anterior.

# Rankings:

Ranking Vectorial: Ranking RSJ: Ranking BM25:



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.