




Moogoo!

 Buscar

Moogoo es una aplicación Web que resuelve el problema de buscar un texto determinado (Query) en un conjunto de documentos de acuerdo al peso(relevancia) de cada palabra en cada documento.

Se calculó la relevancia de las palabras utilizando el Algoritmo de TFIDF(usa la frecuencia de las palabras para determinar que tan relevante son esas palabras en los documentos dados) y luego se calcularon las puntuaciones de equivalencia de la Consulta con respecto a cada documento usando la similitud de cosenos.

Descripción del procedimiento una vez realizada la consulta:

1-Se normaliza el texto de la Consulta.

2-Se itera por cada documento .txt de la carpeta Content, normalizando así su texto.

En cada una de las iteraciones se calcula el TF(Term Frequency) para cada palabra del documento i-th, se construye un vocabulario para el documento i-th y se aumentan los DF(Document Frequency) de las palabras encontradas. Todo esto a través de la función get_frequency.

3-Se procede a calcular los TFIDFs de la Consulta mediante la función query_tfidf.

Esta función calcula los TF(Term Frequency) primeramente, luego transforma en cada palabra los DF(Document Frequency) en IDF(Inverse Document Frequency) mediante la fórmula del IDF(anexo1). Una vez hecho esto, multiplica los TF con los IDF(Evidentemente obteniendo los TF-IDF de cada palabra de la Consulta anexo2).

4-Se llama a la función calculate_cosine para calcular las puntuaciones de equivalencia.

Esta función itera por cada documento de la carpeta Content y hace lo siguiente:

Primer For: (itera por el vocabulario del documento i-th)

a) Convierte el DF de la palabra j-th en el vocabulario en su respectivo IDF mediante la fórmula del IDF(anexo1).

b) Calcula el TF-IDF de cada palabra en el documento(anexo2).

c) Obtiene la sumatoria de los cuadrados de los TF-IDF de cada palabra en el documento(necesario para la fórmula de similitud de cosenos en anexo3).

d) Obtiene la sumatoria de las multiplicaciones de los TF-IDF de las palabras de la query que existan en el documento(necesario para la fórmula de similitud de cosenos en anexo3).

e) Obtiene las palabras que aparecen en la Consulta y el documento i-th para conformar el parámetro snippets.

Segundo For:(itera por cada palabra de la Consulta)

a)Obtiene la sumatoria de los cuadrados de los TF-IDF de cada palabra de la Consulta(necesario para la fórmula de similitud de cosenos en anexo3).

Luego de esto, se calcula el score del documento i-th mediante la fórmula de similitud de cosenos(anexo3).

Se devuelven las puntuaciones de cada documento ordenadas de manera descendente junto con el nombre del documento y su respectivo parámetro snippets.

5- Se convierten estos resultados ordenados a SearchItem y luego se muestran en pantalla.

Estructuras de Datos Utilizadas:

-Diccionarios

-Arrays

-Listas

Implementaciones Adicionales:

1-Se hizo uso del parámetro Suggestion en caso de que se retornen menos de 5 resultados y que al menos una palabra de la Consulta no apareciera en el Corpus. Para esto se hizo uso de la Distancia de Levenshtein(algoritmo usado para saber la menor distancia entre dos cadenas de texto mediante las operaciones de inserción, supresión y reemplazo). Este algoritmo fue implementado usando Programación Dinámica con un arreglo bidimensional.

2- Operador '!' para que la búsqueda no devuelva documentos con las palabras precedidas por este operador. Para esto se hizo uso de una Lista que tomara las palabras precedidas por este operador(método Query) y se excluyeron los documentos conteniendo estas palabras en los resultados(esto en el método calculate_cosine).

3-Operador '^' para que la búsqueda solo devuelva documentos con las palabras precedidas por este operador. Para esto se hizo uso de una Lista que tomara las palabras precedidas por este operador(método Query) y se incluyeron en los resultados solo los documentos que contenían estas palabras.

$$\log \left(\frac{N}{df_i} \right)$$

Anexo1: Calcular IDF con el DF

$$w_{i,j} = tf_{i,j} \times \log \left(\frac{N}{df_i} \right)$$

Anexo2: Calcular el TF-IDF

$$\text{cosine}(\mathbf{v}, \mathbf{w}) = \frac{\mathbf{v} \cdot \mathbf{w}}{|\mathbf{v}| |\mathbf{w}|} = \frac{\sum_{i=1}^N v_i w_i}{\sqrt{\sum_{i=1}^N v_i^2} \sqrt{\sum_{i=1}^N w_i^2}}$$

Anexo3: Calcular la similitud de cosenos con los vectores Doc y Query