Moogle!

Dayan Cabrera Corvo

25 de julio de 2023

Un buscador eficiente.

Facultad Matemática Computación
Universidad de La Habana

Temas a tratar

Moogle! es una aplicación totalmente original cuyo propósito es buscar inteligentemente un texto en un conjunto de documentos. Es una aplicación web, desarrollada con tecnología .NET Core 6.0, específicamente usando Blazor como framework web para la interfaz gráfica. La aplicación está dividida en dos componentes fundamentales: MoogleServer es un servidor web que renderiza la interfaz gráfica y sirve los resultados. MoogleEngine es una biblioteca de clases donde está implementada la lógica del algoritmo de búsqueda.

Modelo matemático

La formulación general del modelo utilizado es TF IDF

$$\begin{cases} TF = n/N \\ IDF = log(1 + D/d) \end{cases} \quad (1)$$

La variable $''n''$ representa la cantidad de apariciones de una palabra en un documento y $''N''$ representa la cantidad de palabras del documento. En el caso del IDF la variable $''D''$ representa la cantidad de documentos existentes y $''d''$ la cantidad de documentos que contienen la palabra en cuestion.
De esta forma al realizar la multiplicacion del TF y el IDF, obtendriamos la relvancia de esa palabra en esos documentos.

Una vez calculado el TDIDF de todos los documentos se procede de la siguiente forma:

- Si $TFIDF = 0$, el documento no tendria relevancia y no se moestraria. equilibrio.
- Si $TFIDF > 0$, el documento se compararia con otros y se mostrarian en orden descendente.

Una ves hecho esto los documentos quedan organizados por orden de relevancia.

La distancia de Levenshtein, distancia de edición o distancia entre palabras es el número mínimo de operaciones requeridas para transformar una cadena de caracteres en otra, se usa ampliamente en teoría de la información y ciencias de la computación. Se entiende por operación, bien una inserción, eliminación o la sustitución de un carácter. Esta distancia recibe ese nombre en honor al científico ruso Vladimir Levenshtein, quien se ocupó de esta distancia en 1965. Es útil en programas que determinan cuán similares son dos cadenas de caracteres, como es el caso de los correctores ortográficos.

- casa cala(sustitucion 's'por 'l')
- cala calla(insercion de 'lénte 'l'y á')
- calla calle(sustitucion de á'por é')

		E	L	E	P	H	A	N	T
	0	1	2	3	4	5	6	7	8
R	1	1	2	3	4	5	6	7	8
E	2	1	2	2	3	4	5	6	7
L	3	2	1	2	3	4	5	6	7
E	4	3	2	1	2	3	4	5	6
V	5	4	3	2	2	3	4	5	6
A	6	5	4	3	3	3	3	4	5
N	7	6	5	4	4	4	4	3	4
T	8	7	6	5	5	5	5	4	3

FIGURA: Ejemplo de distancia de Lvenshtein

Funcionalidades del Moogle

Cuando el programa se ejecuta, en primer lugar se extraen los txt de la ruta asignada, estos textos se procesan y luego a partir de ellos se calcula el TD, el IDF y el TF-IDF para cada palabra diferente en el documento. Una vez procesada toda esta informacion ya las busquedas saldrian mas rapido. Ahora, las busquedas a su vez son procesadas y llevadas a un array que se ira multiplicando por la (Matriz) con los TFIDF de cada texto, luego se le aplicarian los cambios con respecto a los simbolos y esto nos daria el escore de cada texto. En caso de que una palabra de la busqueda no se encuentre en el conjunto de textos, Moogle le enviara una sugerencia, eso aplica a su vez para palabras mal escritas. Cuenta a su vez con caracteres especiales de busqueda: ('!', '^', ' ', '*')

Caracteres especiales

'!'

Delante de una palabra devuelve txt donde esta palabra no puede aparecer.

'^'

Delante de una palabra devuelve txt donde esa palabra tiene que aprecer obligatoriamente.

'*'

Delante de una palabra, le da a esta palabra mas relevancia en la busqueda.

Conclusiones

En resumen Moogle! es un buscador solo de txt basado en un modelo vectorial(TFIDF), se usan ademas otros algoritmos compo la distancia de Levenshtein, creando asi un proyecto considerablemente eficiente.