Moogle!

Pedro Dennis Perez Marquez

25 de julio de 2023

Buscador Moogle.

Facultad Matemática Computación
Universidad de La Habana

Temas a tratar

Moogle! es una innovadora aplicación diseñada para buscar de manera inteligente un texto específico en un conjunto de documentos. Esta aplicación web ha sido desarrollada con tecnología .NET Core 6.0, utilizando Blazor como framework web para la interfaz gráfica. Moogle! se divide en dos componentes principales: MoogleServer, que es un servidor web capaz de renderizar la interfaz gráfica y ofrecer los resultados; y MoogleEngine, una biblioteca de clases que implementa la lógica del algoritmo de búsqueda.

Modelo matemático

La formulación utilizada 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, también conocida como distancia de edición o distancia entre palabras, es una métrica ampliamente utilizada en teoría de la información y ciencias de la computación. Se define como el número mínimo de operaciones necesarias para transformar una cadena de caracteres en otra, donde una operación puede ser una inserción, eliminación o sustitución de un carácter. Esta distancia es especialmente útil en programas que determinan la similitud entre dos cadenas de caracteres, como los correctores ortográficos. La distancia de Levenshtein lleva el nombre del científico ruso Vladimir Levenshtein, quien la propuso por primera vez en 1965. Con su amplia aplicación en la informática, esta métrica ha demostrado ser una herramienta valiosa para la comparación y análisis de cadenas de caracteres.

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

		m	e	i	l	e	n	s	t	e	i	n
	0	1	2	3	4	5	6	7	8	9	10	11
l	1	1	2	3	3	4	5	6	7	8	9	10
e	2	2	1	2	3	3	4	5	6	7	8	9
v	3	3	2	2	3	4	4	5	6	7	8	9
e	4	4	3	3	3	3	4	5	6	6	7	8
n	5	5	4	4	4	4	3	4	5	6	7	7
s	6	6	5	5	5	5	4	3	4	5	6	7
h	7	7	6	6	6	6	5	4	4	5	6	7
t	8	8	7	7	7	7	6	5	4	5	6	7
e	9	9	8	8	8	7	7	6	5	4	5	6
i	10	10	9	8	9	8	8	7	6	5	4	5
n	11	11	10	9	9	9	8	8	7	6	5	4

FIGURA: Anexo1: Distancia de Lvenshtein

Funcionalidades del Moogle

Durante la ejecución del programa, se realiza una extracción inicial de los archivos de texto ubicados en la ruta asignada. Estos textos se procesan y se calcula el TD, IDF y TF-IDF para cada palabra distinta presente en el documento. Este proceso garantiza que, una vez que toda la información ha sido procesada, las búsquedas se realicen de manera más rápida y eficiente. Al realizar una búsqueda, el programa la procesa y la almacena en un array, que, a su vez, se multiplica por la matriz que contiene los TF-IDF de cada texto. Se aplican los cambios correspondientes en relación con los símbolos y, de esta manera, se obtiene el puntaje de cada texto. En el caso de que una palabra de la búsqueda no se encuentre en el conjunto de textos, Moogle! le proporcionará sugerencias relevantes. De manera similar, si se cometió un error de ortografía en la búsqueda, Moogle! detectará automáticamente la palabra correcta y la mostrará en los resultados. Cuenta a su vez con

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 de archivos de texto (.txt) que utiliza un modelo vectorial basado en TF-IDF. Además, incorpora algoritmos como la distancia de Levenshtein para mejorar la eficiencia de la búsqueda. En general, Moogle! es un proyecto altamente eficiente que ofrece una solución efectiva para la búsqueda de información en archivos de texto.