Proyecto Moogle

Abraham Romero Imbert

Facultad de Matemática y Computación
Universidad de la Habana

Julio, 2023

Contenido

1. ¿Qué es Moogle?
2. ¿Cómo funciona?

¿Qué es Moogle?

Figura: Imagen de Moogle!

Es el primer proyecto de Programación orientado a Primer Año de Ciencias de la Computación. Básicamente es una aplicación cuyo propósito es buscar inteligentemente un texto en un conjunto de documentos.

¿Cómo funciona?

Como cualquier buscador busca información en una base de datos a partir de palabras clave o frases que el usuario introduce en la barra de búsqueda. Utiliza algoritmos para encontrar archivos que contienen las palabras clave y las presenta al usuario en una lista de resultados teniendo en cuenta la relevancia de los mismos respecto a la búsqueda realizada Adicionalmente este buscador tiene la capacidad de sugerir al usuario otras posibles búsquedas, en especial al no encontrar resultados en la base de datos

Algoritmo del Programa

Explicación básica del funcionamiento

Debido a que el algoritmo utilizado en mi proyecto es explicado más detalladamente en un informe dentro del mismo proyecto en esta presentación solo se eplicará de manera superficial

Algoritmo del Programa

Explicación básica del funcionamiento

El programa se ejecuta dentro de la clase Moogle y desde la ubicación del programa en la computadora Luego va a la carpeta Content(donde se encuentran los archivos o base de datos entre los cuales se darán los resultados) y analiza cada uno de los documentos para generar un objeto de la clase DiccionarioReferencial el cual almacena los datos necesarios para efectuar el cálculo de la relación TF-IDF (mide la relevancia de una palabra o parte de un documento con el query del usuario)

Luego se genera otro objeto estático de la clase MatrizTFIDF que es la representación en valores de TF-IDF de cada documento y término de la base de datos Este objeto es en sí un array de objetos ¨Vector¨. Cada uno de estos vectores n-dimensionales(n definido por la cantidad de términos diferentes) están representados por valores double definidos por la relevancia(o repetición) de cada término en el documento y en el espacio de la base de datos. El cálculo de TF-IDF se efectúa con la siguiente fórmula:

$$TF \cdot IDF \tag{1}$$

Donde

$$TF_i = \frac{\log_2(Freq(i,j)+1)}{\log_2(L_j + 0{,}001)} \tag{2}$$

Freq(i,j)+1 = Frecuencia del término i en el documento j.
L_j = Número total de términos en el documento j. En este caso se le sumo a este valor 0.001 para evitar indefiniciones en la función

$$IDF_i = log_2(1 + \frac{N_d}{f_i + 1}) \tag{3}$$

N_d = Número total de documentos considerados
f_i = Número de documentos que contienen el término i. En este caso se le sumo a este valor 1 para evitar indefiniciones en la función

Una vez el usuario ha ingresado el query(o cunsulta) la clase Query-class se encarga de analizar dicha entrada y procesarla para después compararla con cada documento y devolver los resultados. Esto se hace mediante la fórmula siguiente que mide la similitud de dos vectores en un espacio n-dimensional:

$$\cos(\theta) = \frac{A \cdot B}{\| A \| \| B \|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}} \quad (4)$$

Dentro de esta misma clase también se elabora un Snippet (o un fragmento del documento donde aparece su parte más parecida al query del usuario), así como una sugerencia de query válido lo más parecido posible al del usuario. Ejemplo:

Figura: Sugerencia en Moogle!

Resumen

De esta manera se ha explicado brevemente el funcionamiento o implementación de mi proyecto Moogle!.Este puede estar sujeto a cambios más adelante que le brinden más funcionaliad o eficiencia. Tenga en cuenta que dentro del mismo proyecto se encuentra un informe que explica más detalladamente cada proceso de ejecución del programa.

Proyecto Moogle

Abraham Romero Imbert

Facultad de Matemática y Computación
Universidad de la Habana

Julio, 2023

Muchas Gracias