Moogle!: Proyecto de Programación I

Cristhian Delgado Garcia

Introducción

Moogle! es un motor de búsqueda simple que usa el Modelo de Espacio Vectorial para encontrar los archivos de texto más relevantes para una consulta. A continuación se describen los conceptos matemáticos y la implementación en C# de este proyecto.

Desarrollo

Se comienza seleccionando el conjunto de documentos con estensión *.txt* en la carpeta *Content* como se puede ver en la figura **??**.

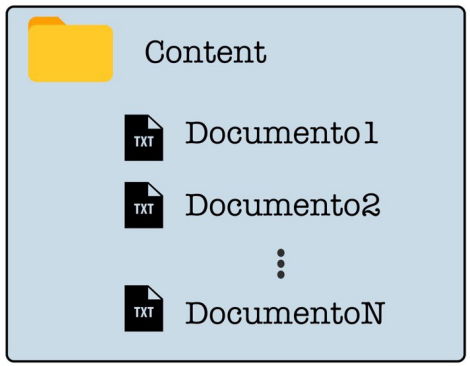

Figura: Carpeta Content

Luego se crea un objeto `documents` de la clase `Documents` con la ruta de ese directorio . Este objeto tiene una propiedad fundamental: `TF_IDF`.La propiedad es un *array* bidimensional que representa la matriz con la medida *tfidf*. Esta medida indica la relevancia de un término en un documento dentro de una colección, y se obtiene multiplicando la frecuencia del término (tf) por el inverso de la frecuencia del documento (idf), que es el logaritmo del cociente entre número de documentos y el número de documentos con el término.Aqui se muestra una posible representación de la matriz TF-IDF:

$$\mathrm{D} = \begin{bmatrix} \mathrm{tfidf}(t_1, d_1) & \mathrm{tfidf}(t_1, d_2) & \cdots & \mathrm{tfidf}(t_1, d_n) \\ \mathrm{tfidf}(t_2, d_1) & \mathrm{tfidf}(t_2, d_2) & \cdots & \mathrm{tfidf}(t_2 d_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{tfidf}(t_m, d_1) & \mathrm{tfidf}(t_m, d_2) & \cdots & \mathrm{tfidf}(t_m, d_n) \end{bmatrix}$$

Como se aprecia el elemento en la posición (i,j) representa el tfidf del i-ésimo término en el j-ésimo documento.

Mientras se procesan los documentos tambien se renderiza la interfaz gráfica como se observa en la figura **??**, en donde el usuario podrá escribir su consulta.

Figura: Interfaz Gráfica

Al escribir el usuario su consulta se llama al método estático `Query` de la clase `Moogle`. Este método crea una instancia `userInp` de la clase `Query` con esa consulta que, igual que el objeto `documents` ,tiene una propiedad `TF_IDF`. Esta propiedad almacena un *array* de `double` con los tfidf de los términos del conjunto de documentos en esta consulta, se puede representar como un vector fila:

$$q = \begin{bmatrix} \text{tfidf}(t_1, q) & \text{tfidf}(t_2, q) & \cdots & \text{tfidf}(t_n, q) \end{bmatrix}$$

Con el método estático `Cos()` de la clase `Vector` se calcula la **similitud coseno** entre el vector query y cada vector documento y se almacena en un array de `double` `scores`. La **similitud coseno** es una medida que dice que tan similares son el vector query y un vector documento, y se calcula como el cociente entre el producto punto de los vectores y el producto de sus respectivos módulos. Por supuesto mientras mayor sea este valor más **relevante** es ese documento para esa query.

Una posible interpretación geométrica para esta medida se puede observar en la figura **??**[1]:

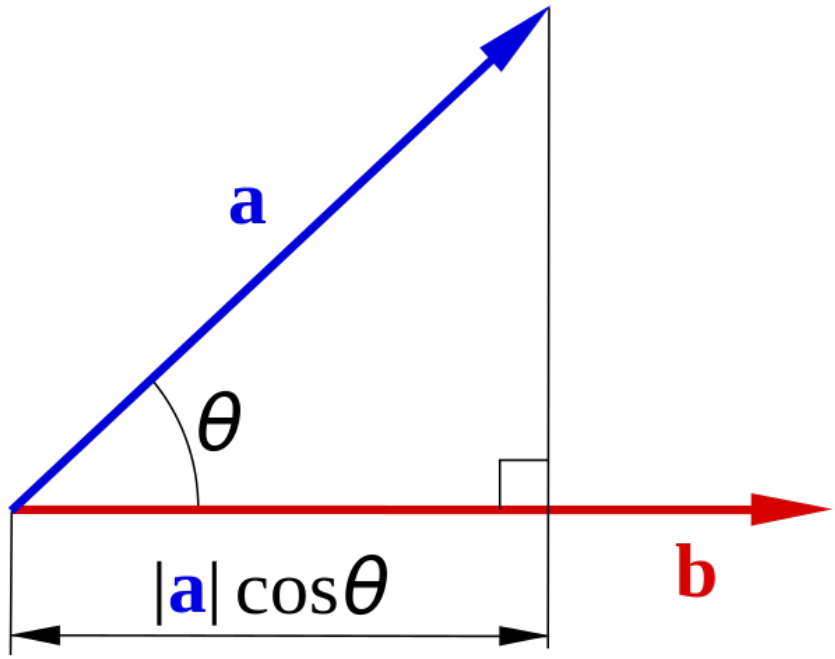

Figura: Similitud coseno

[1] De Svjo - Trabajo propio, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=16746151

Finalmente usando los métodos estáticos `Sort()` y `Reverse()` de la clase *built-in* `Array` se devuelve un objeto `SerachResult` con los documentos ordenados por su relevancia (si tienen alguna relevancia) para esa query y un fragmento de su texto que serán desplegados en la Interfaz Gráfica como se muestra en la figura **??** :

Figura: Resultados de una búsqueda

Conclusiones

Se concluye que **Moogle!** es un proyecto didáctico que ilustra los principios básicos de la recuperación de información y que ofrece una experiencia de búsqueda satisfactoria al usuario.

Bibliografía

- Wikipedia
- Chat Bing
- ChatGPT