Proyecto Moogle!

Jorge Alejandro Echevarría Brunet

Facultad de Matemática y Computación
Universidad de la Habana

¿ Qué es Moogle! ?

Moogle! es una aplicación web **totalmente original** desarrollada con tecnología .NET Core 7.0, específicamente usando Blazor como **framework** para la interfaz gráfica, y en el lenguaje C#, cuyo propósito es buscar inteligentemente un texto en un conjunto de documentos.

Arquitectura Básica

Durante el desarrollo de la aplicación fueron creadas las clases:

- Reader: Identifica los documentos y normaliza sus textos correspondientes.
- TF-IDF: Calcula el peso de los documentos previamente identificados.
- Search: Normaliza la búsqueda (query) introducida y determina los documentos más relevantes.
- Operators: Trabaja con los operadores utilizados en la query, influyendo directamente en el puntaje de los documentos con más relevancia.
- Snippet: Muestra una porción de texto de cada documento relevante y una posible sugerencia al usuario.
- Initialize: Utiliza las funciones de las clases **Reader** y **TF-IDF** a la hora de lanzar la aplicación.

La estructura de datos más importante utilizada son los diccionarios.
Haciendo uso de estos se realizan las operaciones pertinentes para determinar los mejores resultados de acuerdo a la consulta realizada.

Tengamos en cuenta que los principales diccionarios a utilizar durante el Preprocesamiento son:

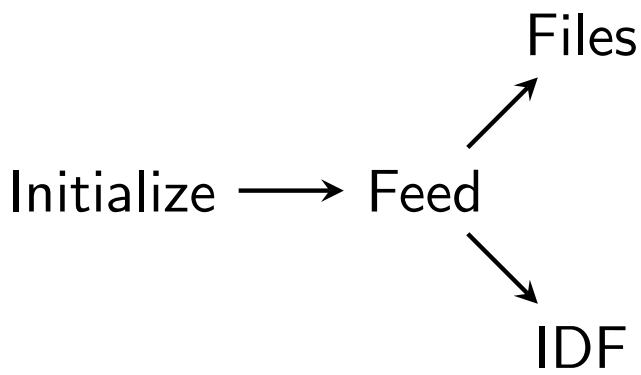
- Files
- IDF

Mientras que durante la Consulta:

- QueryWeight
- Score

Flujo de Datos: Preprocesamiento

Al iniciarse la aplicación se hace un llamado al método **Feed** de la clase **Initialize**, el cual es el encargado de llamar a su vez a los métodos correspondientes de las clases **Reader** y **TF-IDF**, permitiendo que a la hora de realizar la consulta se hayan calculado los pesos de los términos en cada documento, así tambien como la Frecuencia Inversa de Documento (IDF) de cada uno.



La IDF de cada término i en el total de documentos N está dada por la fórmula:

$$IDF_i = \lg_{10}\left(\frac{N}{n_i}\right)$$

Mientras que el cálculo de pesos de cada término i en el documento j está dado por la fórmula:

$$W_{ij} = \frac{freq_{ij}}{max_l\, freq_{lj}} IDF_i$$

Búsqueda

Al realizar la consulta se le aplica una normalización a sus términos y se le asocia a cada uno su peso mediante la función **FeedQueryWeight**, perteneciente a la clase **Search**, para ello se empleó la fórmula:

$$W_{ij} = (\alpha + (1 - \alpha)\frac{freq_{iq}}{max_l freq_{lq}}) \lg_{10} (\frac{N}{n_i})$$

Realizada esta acción se procede a buscar entre los documentos aquellos de mayor relevancia mediante la función **FeedScore**, también perteneciente a la clase **Search**, para determinar la relevancia de cada documento se emplea la fórmula de Similitud de Cosenos, dada por:

$$Rsim(d_j, q) = \frac{\sum_{i=1}^{n} W_{ij} W_{iq}}{\sqrt{\sum_{i=1}^{n} (W_{ij})^2} \sqrt{\sum_{i=1}^{n} (W_{iq})^2}}$$

$$\text{Query} \longrightarrow \text{FeedQueryWeight} \longrightarrow \text{FeedScore} \longrightarrow \text{Best Results}$$

En caso de introducir algún/unos de los operadores:
$[\sim\ !\ \wedge\ *]$, estos influirán en el puntaje final de los resultados encontrados.