Alumno: Carlos Antonio Bresó Sotto

Grupo: C121

Aquí, el proyecto.

Una vez abierta la página de Moogle, este buscador selecciona en la carpeta content los archivos .txt que mejor respondan a la búsqueda insertada. Todo el trabajo esta principalmente en Moogle.cs y Tools.cs, clases donde están todos los métodos que usa el buscador.

Requisitos para que funcione:

Para empezar la carpeta “Content” debe de tener al menos un archivo .txt para que el programa funcione, puse algunos libros allí, pero puedes cambiarlos por cualquier otro cuerpo de búsqueda.

No se debe borrar el “Sinonimos.json” que esta dentro de MoogleEngine. (si desea ver resultados mas cercanos al query puede dejar en comentario la línea 30 de Program.cs y la búsqueda trabajara sin tener en cuenta sinónimos )

Funcionamiento de búsqueda:

A grandes rasgos.

Una vez insertado el query a buscar, se limpia el mismo de signos de puntuación que estorban y se pasa directamente a calcular el TF-IDF teniendo en cuenta el uso de operadores de búsqueda. La búsqueda devuelve un array de objetos tipo SearchItem los cuales están clasificados según que tantas palabras del query tengan, si tienen las palabras mas importantes y si las palabras que contienen están afectadas por algún operador. Todas las búsquedas están acompañadas de una sugerencia de la frase escrita en el query, la cual puedes clicar en la página del buscador para buscar esta sugerencia. En la búsqueda no se tienen en cuenta las palabras demasiado comunes, las cuales son detectadas si su IDF es demasiado bajo (menor que una constante arbitraria). Si por casualidad una palabra poco importante no es desechada por el buscador, sus documentos devueltos aparecen de seguro al fondo.

TF-IDF y operadores:

El TF -IDF no es mas que la multiplicación de la variable TF por IDF. El TF solo tiene en cuenta el texto donde se va a buscar la palabra y es la razón de la cantidad de veces que aparece cierta palabra t en un texto D (TF = t / D). Elegí hacer esta división solo para darle más importancia al archivo si la tasa de aparición de la palabra es alta. El IDF es el valor que tiene una palabra con respecto al cuerpo, es mayor mientras menos documentos la contengan; para una palabra f es el numero de documentos donde aparece y T el total de documentos entonces IDF = T/f.

Operador de cercanía:

Este solo afecta a documentos que contengan las palabras afectadas por el operador. El score se multiplica por una variable que depende de la distancia entre ambas palabras la cual tiene la formula: 1 + (100 / d), donde d es la “distancia entre ambas palabras. Esta variable como se puede apreciar siempre es mayor que 1 así que no afecta negativamente al score.

Demas operadores:

Funcionan, nada interesantes.

Hasta aquí la explicación del buscador. See ya