Informe del Moogle

Daniel Collazo Aldana

Grupo C112

July 24, 2023

Introducción

Este código es una implementación de un motor de búsqueda simple utilizando el modelo TFIDF (Term Frequency-Inverse Document Frequency) para calcular la relevancia de los documentos en función de una consulta de búsqueda.

Funciones

- LeerArchivosDeTexto()
- Normalizar(string documento)
- ObtenerPalabras(string nombreArchivo)
- CalcularVectorTfIdf(string[] palabras, Dictionary¡string, List¡string¿¿ frecuencia, string[] array)
- CalcularSimilitudCoseno(double[] vector1, double[] vector2)
- ObtenerPalabrasQuery(string query)
- CalcularMatrizTfIdf(string[] contenido, string[] nombres)
- CalcularSimilitudQuery(string query, string[] nombres, string[] contenido, double[,] matriz, Dictionary¡string, List¡string¿¿ frecuencia)
- Busqueda(string query)

Funciones (cont.)

- LeerArchivosDeTexto(): lee todos los archivos de texto en una carpeta llamada "Content" y devuelve una tupla con el contenido de los archivos y sus nombres.
- Normalizar(string documento): toma un documento y lo normaliza, convirtiendo todas las palabras en minúsculas y eliminando los caracteres no alfanuméricos.
- ObtenerPalabras(string nombreArchivo): lee el contenido de un archivo y lo normaliza, luego devuelve un array con todas las palabras del archivo.

Funciones (cont.)

- CalcularVectorTfIdf(string[] palabras, Dictionary¡string, List¡string¿¿ frecuencia, string[] array): toma una lista de palabras, una frecuencia de documentos y un array de palabras, y calcula el vector TF-IDF para cada palabra en la lista.
- CalcularSimilitudCoseno(double[] vector1, double[] vector2): calcula la similitud coseno entre dos vectores.
- ObtenerPalabrasQuery(string query): normaliza una consulta de búsqueda y devuelve un array con todas las palabras.

Funciones (cont.)

- CalcularMatrizTfIdf(string[] contenido, string[] nombres): calcula la matriz TF-IDF para todos los documentos en una lista de contenido y nombres de archivo.
- CalcularSimilitudQuery(string query, string[] nombres, string[] contenido, double[,] matriz, Dictionary¡string, List¡string¿¿ frecuencia): calcula la similitud coseno entre una consulta de búsqueda y todos los documentos en una lista de contenido y nombres de archivo.
- Busqueda(string query): es la función principal que realiza una búsqueda de texto. Utiliza las funciones anteriores para calcular la relevancia de los documentos en función de una consulta de búsqueda y devuelve un array de SearchItem, que contiene el nombre del archivo, la similitud coseno, las palabras clave (las palabras que aparecen en el archivo) y las primeras diez palabras del archivo.