Proyecto Final de Sistemas de Recuperación de la Información

Informe

- Daniela Rodríguez Cepero Grupo: C311
- Belsai Arango Hernández Grupo: C311
- Carlos Carret Miranda Grupo: C312

Abstract. Currently Information Retrieval Systems are widely used in various areas, search engines use them constantly. This report describes each stage of the process carried out for the implementation of an Information Retrieval System. Details of tools will be covered used and most important aspects of the code.

Keywords: web, Information Retrieval System

Resumen En la actualidad los Sistemas de Recuperación de Información son muy usados en diversas áreas, los motores de búsqueda los utilizan constantemente. En el presente informe se describe cada etapa del proceso llevado a cabo para la implementación de un Sistema de Recuperación de la Información. Se abordaran los detalles de las herramientas empleadas y aspectos mas importantes del código.

Palabras claves: web, Sistema de Recuperación de Información.

1 Introducción

La información es conocimiento sobre un determinado hecho o circunstancia. La recuperación se refiere a la búsqueda a través de la información almacenada para encontrar información relevante. Ante esto, los sistemas de recuperación de información (SRI) se ocupan de la representación, el almacenamiento, la organización de/y el acceso a los elementos de la información. Los tipos de elementos de información incluyen documentos, páginas web, catálogos en línea, registros estructurados, objetos multimedia, entre otros. La función principal de los SRI son indexar texto y buscar documentos útiles en una colección para darle respuesta a las múltiples consultas de los usuarios.

2 Diseño

Para el funcionamiento del sistema se tiene:

Se tiene una colección de documentos, los cuales son procesados para acceder a su información de manera mas sencilla. El usuario introduce una consulta en forma de texto en el sistema de recuperación de información. Luego, el sistema procesa esta consulta y busca los documentos relevantes para el usuario de la colección que fue procesada. Una vez realizado este proceso, se le devuelven al usuario el título de estos documentos que fueron relevantes.
Utilizaremos uno de los modelos clásicos dentro de la recuperación de información, el modelo booleano para, encontrar la similitud entre los documentos y las consultas realizadas por el usuario.

Escogimos este modelo debido a las ventajas que representa. Modelo simple basado en conjuntos. Fácil de comprender e implementar. Consultas con precisión semántica, aunque no siempre es simple traducir una necesidad informacional a una expresión boolena.

3 Herramientas de desarrollo

Utilizamos el lenguaje de programación python 3.10
Los datos fueron procesados con la ayuda de la librería nltk en la clase Parser. Esta nos ayudó a la separación de los textos de la consulta y los documentos en términos y de ella obtuvimos los stopwords que luego fueron removidos en los textos procesados.
Para almacenar cada documento se utiliza una clase Doc, la cual contiene el id, título y los términos del documento procesado.
Se utiliza la clase boolean model para implementar las propiedades y métodos de este modelo.
A modo resumen la clase recibe la consulta donde está procesada de manera que tiene los términos y los operadores and, or y not que se encuentran de manera explícita en la redacción y que luego todos los términos que no están relacionados entre ellos por un operador se le aplica el operador and.

Uno de los métodos de la clase es el de similitud entre una consulta y un documento.
Para ello el método similitud recibe la consulta y los términos que tiene cada documento a través de un diccionario y lo que hace es que recorre la consulta de izquierda a derecha aplicando los operadores de manera tal que: el operador or es la unión, el operador and es la intersección y el operador not es la diferencia. Lo que hace es que busca los documentos donde se encuentran los términos de la relación actual de manera independiente y establece la operación correspondiente en los documentos obteniendo los que son relevantes y esa información la guarda en una variable que luego trabaja ese resultado con el operador y término que viene en la siguiente iteración.