Proyecto Final de Sistemas de Recuperación de Información. Motor de Búsqueda

Jessy Gigato[1], Laura Tamayo[2], and Yasmin Cisneros[3]

Universidad de La Habana, La Habana, Cuba

Abstract. In our days, search engines have become our best allies and companions in daily use, they help us get to the information we want (in almost all cases) and resolve our most diverse concerns. These are nothing more than mechanisms that organize and distribute the information produced on the network to users who express their doubts from queries in these engines.

Introducción

En nuestros días los motores de búsqueda se han convertido en nuestros mejores aliados y compañeros de uso cotidiano, estos nos ayudan a llegar a la información que queremos (en casi todos los casos) y resolver nuestras mas diversas inquietudes. Estos no son mas que mecanismos que organizan y distribuyen la información producida en la red a los usuarios que expresan sus dudas a partir de consultas en los estos motores. La recuperación de la información se ha convertido en un área sumamente importante en la Ciencia de la Computación ya que es generada diariamente una amplia cantidad de información nueva la cual presenta relevancia y su alcance debería ser importante sin obviar la información precedera.

La recuperación de información es el conjunto de actividades orientadas a facilitar la localización de determinados datos u objetos, y las interrelaciones que estos tienen a su vez con otros. Existen varias disciplinas vinculadas a esta actividad como la lingüística, la documentación o la informática. Con frecuencia, la información responde a qué es algo y que propiedades lo describe, pero tan sólo parte de la información indica cómo se elabora o se desarrolla un proceso. Este tipo de información es básicamente conocimiento. Esta premisa muestra que el conocimiento implica dos cuestiones fundamentales: la existencia de un fin y una relación con otra información de un sistema para lograr un objetivo. En la literatura, la exposición de estas estrategias suele estar vinculada a determinado Sistema de Recuperación. Ya que el desarrollo de estas aplicaciones informáticas surgió como respuesta a la gestión de la sobreabundancia de información actual. La forma en que esta información es almacenada suele ser mediante Bases de Datos y repositorios documentales.

El trabajo presenta como objetivo principal la creación de un Motor de Búsqueda utilizando modelos de recuperación de la información el cual resulte intuitivo al usuario.

Problema Planteado

El Proyecto Final de Sistemas de Recuperación de Información en el curso 2022 consiste en el diseño, implementación, evaluación y análisis de un Sistema de Recuperación de Información. El sistema a desarrollar debe comprender todas las etapas del proceso de recuperación de información. Es decir, desde el procesamiento de la consulta hecha por un usuario, la representación de los documentos y la consulta, el funcionamiento del motor de búsqueda y la obtención de los resultados. No hay limitaciones respecto al Modelo de Recuperación de Información que deben emplear, puede ser cualquiera de los clásicos o alguno de los alternativos, siempre atendiendo a las características de cada uno y su adecuación al escenario en el que se aplicarán.

Desarrollo

Diseño del Sistema:

En el campo de la recuperación de la información se tienen varios modelos clásicos para la recuperación de la misma. Estos serian:

- Modelo Booleano
- Modelo Vectorial
- Modelo Probabilístico

Para la creación de nuestro Motor de Bus queda nos centraremos en la utilización del modelo Vectorial

Arquitectura del Proyecto

El Motor de Búsqueda pasara por una serie de procesos para retornarle al usuario una respuesta:

- Se le entra una consulta
- Realiza el proceso de búsqueda en el corpus de documentos que se tiene
- Retorna una lista de documentos con las coincidencias y/o respuestas mas acertadas

Implementación

Cada uno de los procesos vistos anteriormente constara de una serie de pasos los cuales constituirían el pipeline (flujo) de la aplicación.

Parte de este pipeline sería:
- tokenizar la entrada
- limpiar dichos tokens
- eliminar los stopwords

Representación Como estamos utilizando el modelo vectorial entonces los documentos son representados mediante vectores, cuya representación es obtenida mediante la tokenización de la consulta.

$$w_{ij} = fract f_{ij} \max_k tf_{kj} idf_j \tag{1}$$

Ecuación 1. Representación vista en clases

Donde:
w_{ij} - Representa el peso asociado al termino i en el documento j
tf_{ij} - Número de veces que se repite el término i en el documento j
idf_j - Frecuencia inversa del documento j en la colección de documentos calculada por $\log \frac{N}{n_j}$
n_j - cantidad de documentos en donde aparece el término j.

La consulta presenta la misma representación que los documentos solamente que se le aplica una formula diferente:

$$(\alpha idf_i + (1-\alpha)\frac{tf_{ij}}{\max_k tf_{kj}})idf_i \tag{2}$$

Siendo α el factor de suavizado

Finalmente para realizar un ranking de los documentos se calcula la similitud entre los documentos y la consulta y son ordenados de mayor a menor y asi se devolveran.

$$sim(q, d_i) = \cos(q, d_i) = \frac{q \cdot d_i}{||q||||d_i||} \tag{3}$$

Ecuación 3. Función de similitud

Las principales bibliotecas de Python utilizadas para el proyecto fueron sklearn (utilizada para preprocesar el corpus de documentos) y nltk (usado en la tokenizacion y limpieza de la consulta)

Visual

La aplicación visual fue creada utilizando PyQt.

Conclusiones

(Trabajando)