ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE INGENERIA EN
SISTEMAS INGENERIA EN CIENCIAS
DE LA COMPUTACIÓN

RECUPERACIÓN DE LA INFORMACIÓN

TEMA: Proyecto de Primer Bimestre

INTEGRANTES:

María Fernanda Arias

Andrés Jiménez

Índice

Introducción

Dentro del presente informe se presenta el diseño, programación y ejecución de un Sistema de Recuperación de Información (SRI), para lo cual se utilizó el corpus **"beir/cqadupstack/webmasters"** recuperado de la página web https://ir-datasets.com/, dentro del mismo se trabajó en diferentes procesos necesarios para el tratamiento de los datos como la recuperación del corpus, el preprocesamiento, creación de un vocabulario, generación del índice invertido y trabaja con los modelos de recuperación TF-IDF y BM25 para en base a una query ingresada por el usuario, muestra un ranking de documentos relevantes ordenados por su relevancia y muestra parámetros de evaluación para observar que tan efectivo es el Sistema de Recuperación.

Descripción del corpus utilizado.

Como se mencionó previamente, el corpus utilizado para la ejecución del Sistema fue **"beir/cqadupstack/webmasters",** que forma parte del grupo BEIR (Benchmarking IR), la cual es una colección de 18 datasets orientados a evaluar a los modelos de recuperación en múltiples dominios.

Nuestro caso específico, el corpus es traído de CQADupStack que es un dataset para identificar preguntas duplicadas dentro de diferentes foros, para nuestro sistema se utilizó el sub-dataset “webmaters” que corresponde al foro Webmasters, dentro de este corpus se tiene aproximadamente 17000 documentos y 506 consultas las cuales se basan en preguntas dentro de aquel foro.

Explicación de las decisiones de diseño.

Para el diseño de la aplicación se tomaron en cuenta las siguientes recomendaciones generales:

Preprocesamiento

- Se utiliza regexp_tokenize para separar palabras alfanuméricas.
- Se eliminan palabras vacías con stopwords.words('english').
- Se aplica lematización mediante WordNetLemmatizer.

- Se combinan los tokens del título y el texto del documento para el cálculo de similitudes.

Representación de Documentos

- Para TF-IDF se utiliza TfidfVectorizer de scikit-learn, transformando los documentos en vectores.
- Para BM25 se implementa la fórmula clásica desde cero con factores k1 = 1.5 y b = 0.75.

Índice Invertido

- Se construye manualmente un índice invertido para la implementación de BM25.
- Se almacena un diccionario con la frecuencia de términos por documento.

Evaluación

- Se implementa evaluación automática con métricas como Precisión, Recall y MAP.
- Se emplea el archivo QRELS para definir la relevancia esperada.

Interfaz

Se pensó en una interfaz en línea de comandos (CLI) donde el usuario puede introducir consultas y ver resultados tanto para TF-IDF como para BM25.

Todas estas decisiones fueron tomadas considerando la facilidad de ejecución del código, el rendimiento general de la aplicación, la facilidad del sistema para que el usuario pueda utilizar el Sistema de Recuperación, y medidas de optimización para que el tiempo de respuesta de búsqueda sea en menor tiempo a otros sistemas.

Ejemplos de consultas y resultados.

Para la ejecución del sistema se realizaron diferentes pruebas usando una query ingresada por el usuario mediante la consola, y en base a esa consulta, el sistema realizaba todo el proceso de pre procesamiento y búsqueda, para finalmente mostrar los resultados esperados.

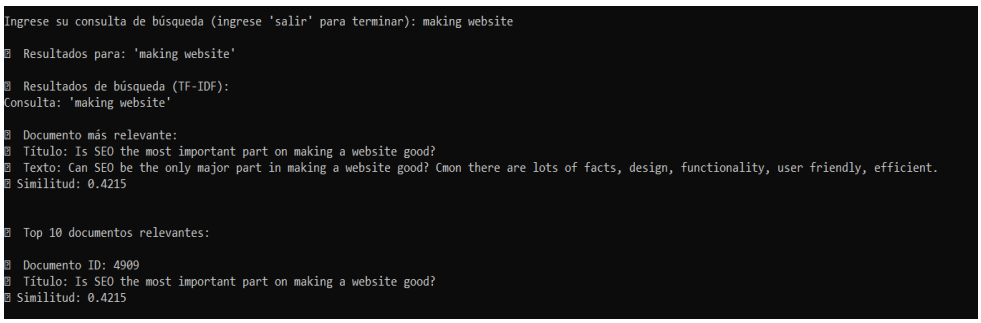

Figura 1. Ejemplo de consulta

Como se observa, la consulta ingresada por el usuario para este ejemplo fue “making website”, el programa realiza el proceso de preprocesar la query, y con el resultado

ejecuta la búsqueda tanto por TF-IDF como por BM25, para luego mostrar los primeros 5 documentos más relevantes ordenados por su relevancia, además de mostrar el título y texto del primero documento en el ranking de relevancia.

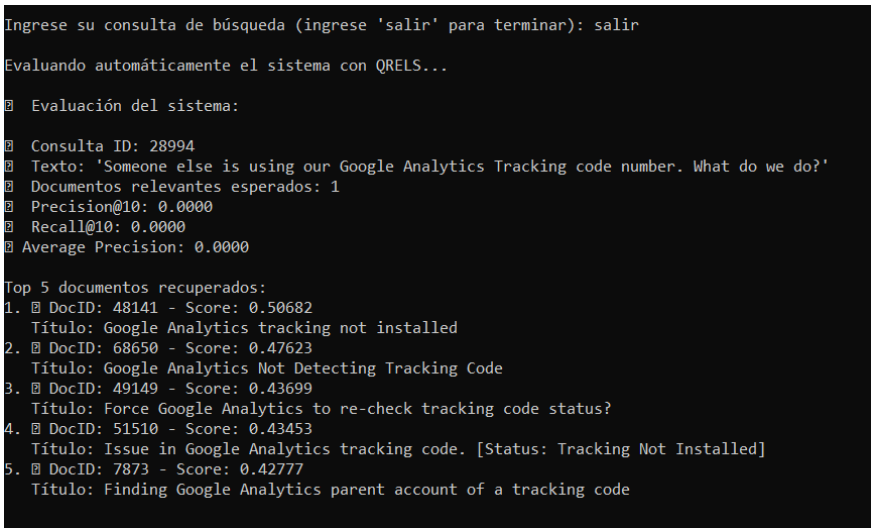

Figura 2. Evaluación del sistema usando qrels.

También se observa que una vez que se coloca la palabra “salida”, el programa hará una evaluación del sistema automática para lo cual utilizara las qrels y queries dadas en el mismo corpus, proporcionando de igual manera los documentos más relevantes ordenados en un ranking, además nos dará el resultado de la evaluación medidos en precisión y el recall, que más adelante se explicara su importancia dentro de la evaluación del sistema.

Análisis de métricas de evaluación.

Para evaluar el rendimiento del sistema de recuperación de información, se emplearon las siguientes métricas: precisión, recall y precisión promedio, calculadas individualmente por consulta. Además, se utilizó la precisión media promedio (MAP) como métrica agregada global.

- Precisión: mide la proporción de documentos relevantes dentro del conjunto de documentos recuperados. Esto es útil para ver qué tan preprocesado está el conjunto de resultados entregados al usuario.
- Recall: mide la proporción de documentos relevantes dentro del conjunto de documentos relevantes que fueron efectivamente recuperados por el sistema.
- Precisión promedio: Evalúa el rendimiento del sistema teniendo en cuenta la posición de los documentos relevantes dentro del ranking. Si es un número alto indica que los documentos relevantes aparecen en las primeras posiciones.
- Mean Average Precision (MAP): Calcula el promedio de los valores AP de todas las consultas de prueba, ofreciendo una medida general del sistema en distintos escenarios de búsqueda.

Durante la ejecución del sistema, se utilizó un archivo “test.tsv” que contiene consultas y documentos relevantes asociados a cada una. El sistema evalúa automáticamente cada

consulta, recupera los documentos más relevantes utilizando la similitud de coseno sobre representaciones TF-IDF, y luego compara los resultados recuperados con el conjunto de documentos relevantes para calcular las métricas mencionadas.

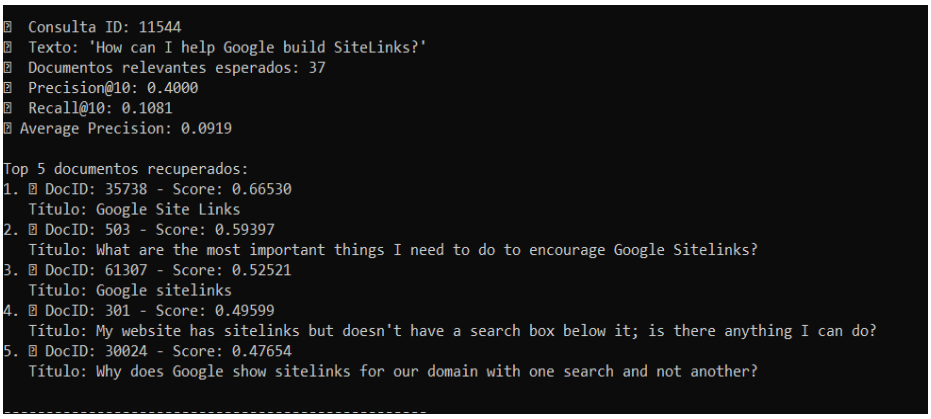

Figura 3. Muestra de documentos relevantes para las qrels.

Como se observa en la imagen, el programa recupera los documentos más relevantes para cada consulta manejando la variable (query-id), y a partir de ahí calcula la precisión, el recall y promedio de precisión (AP) que posteriormente nos servirán para medir el Mean Average Precision (MAP) de todo el sistema.

De esta forma se probó el funcionamiento del Sistema de Recuperación, observando que tanto el corpus como la query ingresada por el usuario siguen los pasos de pre procesamiento y luego de búsqueda tanto para TF-IDF y BM25, además que el sistema se evalúa de forma automatizada usando las qrels proporcionadas por el corpus.

Conclusiones

- Se construyó un sistema de recuperación efectivo con dos métodos de ranking: TF-IDF y BM25.
- El sistema es capaz de responder consultas en línea, evaluar su rendimiento y mostrar documentos originales relevantes.
- Las métricas nos ayudan a verificar la efectividad del sistema, el cual puede ser mejorado en base a los resultados que nos da la Recuperación de la Información.

Video Demostrativo:

Video_Demostrativo.mp4