ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE INGENERIA DE SISTEMAS

INGENERIA EN CIENCIAS DE LA COMPUTACIÓN

RECUPERACIÓN DE LA INFORMACIÓN

SISTEMA DE RECUPERACION DE INFORMACION

PROYECTO BIMESTRAL

INTEGRANTES:

DARLIN ANACICHA

MICHAEL PERUGACHI

1. Introducción

En el presente informe se expone y describe el diseño, la implementación y la evaluación de un Sistema de Recuperación de Información. El objetivo principal del proyecto fue comparar la efectividad de varios modelos de ranking: Similitud de Jaccard, Similitud de coseno utilizando TD-IDF y modelo BM25, mediante la ejecución de consultas de texto libre, además de poder evaluar la calidad de los resultados obtenidos mediante métricas estándar como precisión, recall y MAP.

2. Descripción del corpus utilizado

El corpus utilizado para la construcción y evaluación del sistema es **"beir/cqadupstack/webmasters"** recuperado de la pagina web https://ir-datasets.com/beir.html. Este corpus pertenece a la colección BEIR (Benchmarking IR), una de las bases más usadas para evaluación de Recuperación de Información.
Específicamente el corpus **BEIR – CQAdupstack Webmasters** es un dataset especializado en administración de sitios web, extraído de Webmasters StackExchange. Contiene miles de preguntas técnicas que sirven como documentos (17,000), así como un conjunto de consultas reales (506) y relevancias asociadas (qrels). El contenido del corpus abarca servidores, DNS, PHP, JavaScript, WordPress, seguridad y hosting.

3. Explicación de las decisiones de diseño

Para el diseño del Sistema de Recuperación de Información, se tomaron las siguientes decisiones:
Preprocesamiento

- Tokenización y Normalización: Se utiliza el módulo re para filtrar caracteres no alfabéticos y se convierte el texto a minúsculas (text.lower() y re.sub()).
- Filtrado de Palabras Vacías: Se eliminan las *stop words* con nltk.corpus.stopwords.words('english').
- Reducción Morfológica: Se aplica Stemming mediante PorterStemmer. Esta técnica se eligió para aumentar el *Recall*

Modelos de Ranking

- Modelo Similitud Jaccard: Se implementó para establecer un rendimiento mínimo. Su diseño simple solo mide la superposición de conjuntos, sin ponderar la importancia de los términos.
- Modelo TF-IDF con Similitud Coseno: Se implementó el Modelo Vectorial Clásico para introducir la ponderación de términos.
- Modelo BM25: Se eligió Okapi BM25 como el modelo avanzado. Se implementó la fórmula clásica utilizando los parámetros predefinidos $k1=1.2$ y $b=0.75$.

Indice Invertido

- Construcción Manual del Índice: Se utiliza "defaultdict(dict)" para construir el índice invertido. Esta decisión de diseño permite un acceso rápido a la lista de documentos para cualquier término de la consulta.
- Almacenamiento de Frecuencias: Se almacena la Frecuencia de Término (TF) por documento. Esto como requisito fundamental para el cálculo de los pesos en TF-IDF y BM25.

Evaluación

- Métricas Estándar: Se implementó evaluación automática utilizando Precisión@K, Recall@K, y Mean Average Precision (MAP).
- Uso del Archivo QRELS: La dependencia en el *ground truth* (qrels_df) permite medir el rendimiento de los modelos en función de las relaciones de relevancia definidas por expertos.

Interfaz

- Interfaz Interactiva (ipywidgets): Se eligió una interfaz basada en *widgets* de Jupyter Notebook en lugar de una línea de comandos (CLI). Esta decisión permite al usuario introducir consultas y ver los resultados de los tres modelos de manera simultánea y en tiempo real.

4. Ejemplos de consultas y resultados

Para la evaluación del sistema se realizaron múltiples ejecuciones y pruebas con consultas ingresadas por el usuario mediante la interfaz, todas estas consultas pasaron por el proceso de preprocesamiento y modelos de ranking dándonos los siguientes resultados:

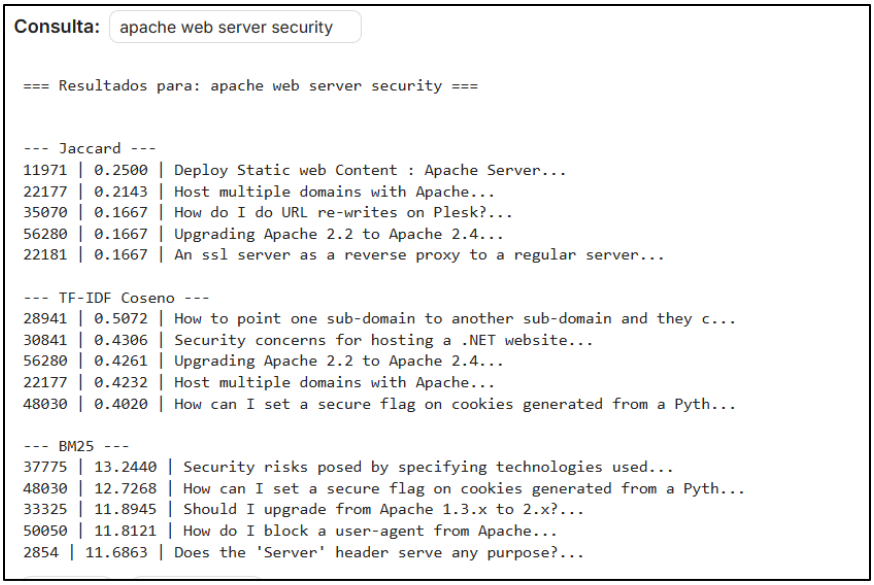

Figura 1. Consulta 1 “apache web server security”

Como se observa en la figura 1, al ingresar la consulta el sistema realiza todo el sistema de preprocesamiento y ranking con los modelos y nos da como resultado los primeros 5 documentos más relevantes y ordenados según el ranking de cada modelo. Adicionalmente, presenta el ID del documento, puntuación de relevancia y el título del documento.

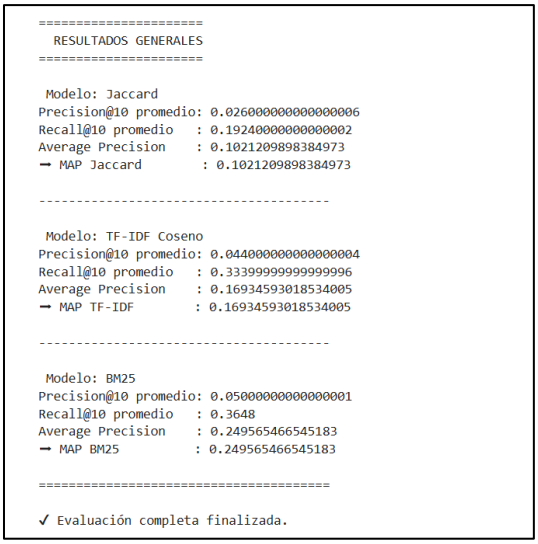

Figura 2. Desempeño de los modelos del sistema.

Como se observa en la figura 2, se nos muestran los resultados de las métricas precisión, recall y map promediadas a lo largo del subconjunto de consultas. Se nos presenta que Jaccard es el más simple y tiene el peor rendimiento, TF-IDF con Similitud Coseno demuestra una mejora sustancial al introducir la ponderación de términos y BM25 es el ganador ya que su capacidad para manejar la longitud de los documentos y evitar que

términos muy frecuentes dominen las puntuaciones resulta en un orden de relevancia superior.

5. Análisis de métricas de evaluación

Para analizar el desempeño del sistema de recuperación de información, se aplicaron diversas métricas: precisión, recall y precisión promedio, todas calculadas de manera individual para cada consulta. Asimismo, se empleó la precisión media promedio (MAP) como medida global resumida del rendimiento general del sistema.

- Precisión: indica qué proporción de los documentos recuperados realmente es relevante. Esta métrica permite evaluar la calidad del conjunto de resultados que se le presentan al usuario.
- Recall: representa la fracción de documentos relevantes que el sistema logró recuperar sobre el total de documentos relevantes existentes.
- Precisión promedio: considera la posición en la que se ubican los documentos relevantes dentro del ranking, de modo que valores altos reflejan que estos aparecen en los primeros lugares.
- Mean Average Precision (MAP): corresponde al promedio de las precisiones promedio obtenidas para todas las consultas de prueba, proporcionando una visión global del rendimiento del sistema en distintos escenarios de búsqueda.

6. Conclusiones

- Se logro construir un sistema de recuperación con 3 modelos de ranking distintos: Similitud de Jaccard, Similitud de coseno utilizando TD-IDF y modelo BM25
- Los resultados de la evaluación demuestran categóricamente que el modelo Okapi BM25 es la opción más efectiva para este corpus. Su capacidad para ofrecer el MAP más alto confirma que es el mejor modelo para ordenar los resultados de búsqueda por relevancia.
- El sistema permite ejecutar consultas en tiempo real, analizar su desempeño y visualizar los documentos relevantes recuperados, además, las métricas empleadas facilitan la evaluación de su efectividad y sirven como guía para identificar oportunidades de mejora basadas en los resultados obtenidos durante el proceso de recuperación de información.

7. Video demostrativo

Link video: https://drive.google.com/file/d/1vb31FXZfAVGOH0E-TzzRmaivXCF67hHt/view?usp=sharing