



ESCUELA POLITECNICA NACIONAL

Facultad de Ingeniería de Sistemas

*Ciencias de la Computación*

Recuperación de la Información

PROYECTO PRIMER BIMESTRE

NOMBRES:

Angel Falcón

Inti Poaquiza

GRUPO:

GR1CC\_2025-2

PROFESOR:

Ivan Carrera

# Sistema de Recuperación de Información

## Objetivos

Diseñar e implementar un sistema de recuperación de información que indexe un conjunto de documentos en texto plano y permita ejecutar consultas de texto libre utilizando el modelo vectorial con vectores binarios y ponderacion TF-IDF, y el modelo probabilistico BM25. El sistema debe permitir evaluar la calidad de los resultados utilizando metricas estandar como precision y recall.

## Introducción

En este proyecto se desarrolló un motor de búsqueda que utiliza dos enfoques complementarios: TF-IDF y BM25. Ambos métodos permiten identificar qué documentos responden mejor a las necesidades de búsqueda de los usuarios. El sistema realiza varias tareas en secuencia:

- Limpia y organiza el texto para trabajar con él más fácilmente, eliminando elementos innecesarios y separando las palabras de forma adecuada.
- Construye una estructura de datos que facilita la búsqueda rápida de términos dentro de los documentos.
- Evalúa qué tan bien funciona el sistema mediante diferentes métricas que miden aspectos como la precisión de los resultados, la capacidad de encontrar información relevante y el rendimiento promedio en múltiples consultas.

La herramienta incluye una interfaz que permite probar ambos algoritmos de forma práctica, comparar sus resultados y observar cuál se comporta mejor según el tipo de búsqueda realizada.

## Descripción del corpus utilizado

Para este proyecto se seleccionó el corpus de noticias de la BBC (BBC News Dataset), que contiene más de 50,000 artículos periodísticos en inglés recopilados de diferentes secciones informativas. Este conjunto de datos presenta características ideales para evaluar sistemas de recuperación, los textos varían considerablemente en longitud, desde noticias breves de último momento hasta artículos de análisis extensos; el vocabulario es rico y diverso, abarcando terminología técnica de diferentes ámbitos junto con lenguaje

periodístico general. El corpus se encuentra estructurado en formato CSV, donde cada registro incluye el título, la categoría temática y la descripción completa del artículo. Para efectos de este sistema se utilizó específicamente el campo de descripción, que contiene el contenido principal de cada noticia y proporciona suficiente información textual para aplicar los algoritmos de ranking sin sobrecargar innecesariamente el procesamiento con metadatos adicionales.

## Explicación de decisiones de diseño

El diseño del sistema se fundamentó en tres decisiones principales: Primero, se implementó un preprocesamiento en cadena donde cada documento atraviesa secuencialmente tokenización, normalización, eliminación de stopwords y stemming. Se eligió stemming con Porter en lugar de lematización porque es computacionalmente más rápido y suficiente para recuperación de información. La eliminación de caracteres no alfabéticos responde a que en noticias los números y símbolos suelen ser contextuales y no discriminativos. Cada etapa se implementó como función independiente para facilitar modificaciones sin afectar el resto del pipeline.

Segundo, la arquitectura modular separa claramente carga de datos, preprocesamiento, indexación, búsqueda y evaluación. Una decisión técnica importante fue construir el índice invertido manualmente con diccionarios anidados aunque las bibliotecas ya lo abstraen, porque esto proporciona control total sobre la estructura y permite entender exactamente cómo se organizan términos y frecuencias. Las interfaces de búsqueda se estandarizaron para que TF-IDF, BM25 y Jaccard reciban consultas procesadas idénticamente y devuelvan resultados en el mismo formato (lista de tuplas doc\_id-score), permitiendo comparaciones directas.

Tercero, se priorizó un diseño comparativo que permite evaluar múltiples modelos con las mismas consultas. Se implementaron qrels (documentos relevantes predefinidos) para cinco consultas de prueba, calculando precisión, recall y MAP de cada modelo. La visualización con tabulate facilita identificar patrones de comportamiento: por ejemplo, detectar si BM25 supera consistentemente a TF-IDF o si algún modelo falla en consultas específicas. Esta capacidad de análisis convierte al sistema en herramienta experimental además de aplicación práctica. Finalmente, se usaron bibliotecas especializadas (NLTK, scikit-learn, rank-bm25) para operaciones estándar en lugar de reimplementarlas. Esto garantiza corrección y eficiencia en los componentes base mientras se mantiene control sobre los aspectos críticos del flujo de procesamiento.

## Ejemplos de consultas y resultados

CONSULTAS DE EVALUACIÓN (QUERIES) Y DOCUMENTOS RELEVANTES (QRELS)		
ID Consulta	Texto de la Consulta	IDs Documentos Relevantes (QRELS)
Q1	Ukraine's youngest cabinet minister	13, 36037, 13511, 7629, 14249
Q2	Russian gymnast Ivan Kuliak is being investigated	11, 78, 53, 104, 3030
Q3	The Ukrainian president says the country will	0, 5527, 2648, 12267, 1372
Q4	TikTok suspends live streaming	7, 31437, 2474, 15057, 23740
Q5	The FIA wants to limit the ways its	42109, 19020, 31481, 31311, 28727

=====			
--- EVALUANDO MODELO: TF-IDF ---			
...			
Consulta ID	Precisión	Exhaustividad (Recall)	AP
Q1	0.0027	1.0000	0.811111
Q2	0.0033	0.8000	0.8
Q3	0.0006	1.0000	0.412844
Q4	0.0053	1.0000	0.711667
Q5	0.0044	1.0000	0.605455
--- MAP ---			0.6682

=====			
--- EVALUANDO MODELO: BM25 ---			
...			
Consulta ID	Precisión	Exhaustividad (Recall)	AP
Q1	0.0027	1.0000	0.588095
Q2	0.0033	0.8000	0.8
Q3	0.0006	1.0000	0.521775
Q4	0.0053	1.0000	0.555556
Q5	0.0044	1.0000	0.764286
--- MAP ---			0.6459

=====			
--- EVALUANDO MODELO: Jaccard ---			
Consulta ID	Precisión	Exhaustividad (Recall)	AP
Q1	0.0027	1.0000	0.354312
Q2	0.0033	0.8000	0.8
Q3	0.0006	1.0000	0.667619
Q4	0.0053	1.0000	0.57972
Q5	0.0044	1.0000	0.703333
--- MAP ---			0.621

## Análisis de métricas de evaluación

El análisis se centra en tres métricas fundamentales: precisión, exhaustividad (recall) y promedio de precisión media (MAP), siendo este último el indicador más relevante para evaluar sistemas de ranking.

### Precisión

Los valores de precisión son consistentemente bajos en todos los modelos (entre 0.0006 y 0.0053), lo cual no representa necesariamente un problema sino una consecuencia directa del tamaño del corpus. Con más de 50,000 documentos y solo 5 documentos relevantes por consulta, incluso recuperar todos los documentos relevantes resulta en precisiones aparentemente bajas. Por ejemplo, en Q4 donde la precisión es 0.0053, esto significa que se recuperaron aproximadamente 943 documentos en total, de los cuales 5 eran relevantes. Lo importante aquí no es el valor absoluto de precisión sino la capacidad del modelo de posicionar esos 5 documentos relevantes en las primeras posiciones del ranking, que es lo que captura el Average Precision.

### Exhaustividad (Recall)

Los tres modelos demostraron una exhaustividad excelente, recuperando todos los documentos relevantes ( $\text{recall} = 1.0$ ) en cuatro de las cinco consultas. La única excepción fue la consulta Q2, donde los tres modelos alcanzaron únicamente 0.8 de recall, indicando que uno de los cinco documentos relevantes no fue recuperado por ningún algoritmo. Este patrón uniforme sugiere que el problema no radica en las capacidades de los modelos sino posiblemente en las características específicas de esa consulta o del documento faltante, que podría contener una terminología suficientemente diferente como para no coincidir con los términos procesados de la consulta tras aplicar stemming y normalización.

## Rendimiento Global

En términos de MAP, que representa la métrica más comprehensiva al considerar tanto la capacidad de encontrar documentos relevantes como su posicionamiento en el ranking, TF-IDF obtuvo el mejor desempeño con 0.6682, seguido por BM25 con 0.6459 y Jaccard con 0.621. Esta diferencia del 2.23% entre TF-IDF y BM25 es relativamente pequeña pero consistente, sugiriendo que ambos modelos son comparables en esta colección específica. La diferencia más significativa se observa con Jaccard, que quedó aproximadamente 4.7% por debajo de TF-IDF, evidenciando las limitaciones de un enfoque puramente basado en coincidencia de conjuntos sin considerar pesos de términos.

## Bibliografía

- [1] Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press. <https://nlp.stanford.edu/IR-book/>
- [2] Robertson, S., & Zaragoza, H. (2009). The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends in Information Retrieval*, 3(4), 333-389. <https://doi.org/10.1561/1500000019>
- [3] Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media. <https://www.nltk.org/book/>