Proyecto de 1er Bimestre
Sistema de Recuperación de Información

Nombres: Fabián Simbaña, Brayan Ortíz

1. Objetivo

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 ponderación TF-IDF, y el modelo probabilístico BM25. El sistema debe permitir evaluar la calidad de los resultados utilizando métricas estándar como *precisión* y *recall*.

2. Corpus

Para la implementación y evaluación del sistema de recuperación de información, se seleccionó la colección **Cranfield,** es un corpus de dominio específico centrado en Ingeniería Aeronáutica y Aerodinámica.

Características Estadísticas:

- **Volumen de Documentos:** 1,400 resúmenes (abstracts) de artículos científicos.
- **Volumen de Consultas:** 225 consultas en lenguaje natural formuladas por expertos en la materia.
- **Juicios de Relevancia (Ground Truth):** El dataset incluye el archivo cranqrel, que contiene aproximadamente 5,285 pares validados de (consulta, documento relevante). Esto proporciona un promedio de ~23 documentos relevantes por consulta, permitiendo una evaluación estadística robusta.

Se optó por Cranfield sobre otras colecciones (como CISI) debido a la precisión de su terminología técnica. Al ser un dominio de ingeniería, los términos como *"número de Mach"*, *"capa límite"* o *"viscosidad"* son discriminantes claros, lo que permite evaluar con mayor fidelidad la eficacia de los modelos vectoriales y probabilísticos.

3. Diseño

Preprocesamiento.- Se identificó que el preprocesamiento estándar (eliminación de signos de puntuación) era perjudicial para textos de ingeniería.

- **Preservación de Guiones:** Se implementó una expresión regular de limpieza para conservar los guiones (-). Términos como steady-flow (flujo constante) o quasi-linear (cuasi-lineal) poseen un significado físico único que se pierde si se separan en unigramas.
- **Stopwords de Dominio:** Además de las palabras vacías del inglés (preposiciones, artículos), se diseñó una **"Lista de Ruido de Ingeniería"**. Se eliminaron términos metodológicos omnipresentes en papers (*calculate, measure, obtain, result,*

paper, report) que no aportan valor semántico para la recuperación, permitiendo que el sistema priorice sustantivos técnicos (*wing, pressure, flow*).

Modelos.- Se implementaron tres modelos matemáticos utilizando Python puro (NumPy/Pandas) para evitar el uso de librerías de caja negra:

1. **Modelo Jaccard:** Utilizado como línea base (baseline) para demostrar las limitaciones de los enfoques binarios que ignoran la frecuencia de los términos.
2. **Modelo Vectorial (TF-IDF):** Se implementó con **Suavizado Logarítmico**. Esto fue necesario porque en textos técnicos la repetición de una palabra no implica una relevancia linealmente proporcional.
3. **Modelo Probabilístico (BM25):** Se seleccionó como el modelo principal. Tras realizar pruebas empíricas, se calibraron sus hiperparámetros a **k_1=1.8** y **b=0.9**.
    - *k_1=1.8:* Un valor alto de saturación para premiar la recurrencia de términos técnicos específicos.
    - *b=0.9:* Una penalización alta por longitud, asumiendo que en resúmenes cortos (abstracts), la longitud excesiva suele correlacionarse con divagación o ruido.

Ejemplo de transformación de texto crudo a tokens procesados.

1. TEXTO ORIGINAL (Raw Input):

'experimental investigation of the aerodynamics of a wing in a slipstream . an experimental study of a wing in a propeller slipstream was made in order...'

2. LIMPIEZA (Regex + Guiones preservados):

'experimental investigation of the aerodynamics of a wing in a slipstream an experimental study of a wing in a propeller slipstream was made in order...'

3. TOKENIZACIÓN + STEMMING (Salida del Índice):

[Tokens]: ['aerodynam', 'wing', 'slipstream', 'wing', 'propel', 'slipstream', 'made', 'order', 'spanwis', 'distribut', 'lift', 'increas', 'due', 'slipstream', 'differ']...

Total Tokens: 64 (Originales: 145)

4. Consultas y Resultados

Para validar cualitativamente el sistema, se realizaron pruebas con consultas técnicas complejas. A continuación, se presenta un caso de estudio con la consulta:

"boundary layer flow separation" (Separación del flujo de la capa límite)

Tabla de Resultados (Top-1 Recuperado por Modelo):

Modelo	Doc ID	Score	Fragmento del Documento Recuperado	Análisis
Jaccard	3	0.2308	*"...the boundary layer in simple shear flow..."*	Recupera un documento relevante pero con un score bajo, incapaz de distinguirlo fuertemente de otros documentos con vocabulario compartido.
TF-IDF	358	0.4248	*"...on the model of the free shock separation..."*	Identifica correctamente la relevancia basándose en la rareza de los términos "shock" y "separation".
BM25	**358**	**11.08**	*"...on the model of the free shock separation..."*	Recupera el mismo documento que TF-IDF pero con un margen de confianza (score) significativamente mayor, discriminando mejor el ruido.

Otro caso de prueba fue la CONSULTA ID 20 para el top 10:

'has anyone formally determined the influence of joule heating, produced by the induced current, ...'

Documentos Relevantes Totales en QRELS: 12

```
>> MODELO 1: JACCARD (Binario) (Top 10):
   1. [Doc 407] ✅ ACERTO | Score: 0.1429 | ...
   2. [Doc 500] ✅ ACERTO | Score: 0.1250 | ...
   3. [Doc 268] ✅ ACERTO | Score: 0.1020 | ...
   4. [Doc 963] ❌ FALLO | Score: 0.0968 | ...
   5. [Doc 269] ✅ ACERTO | Score: 0.0930 | ...
   6. [Doc 450] ❌ FALLO | Score: 0.0882 | ...
   7. [Doc 1158] ❌ FALLO | Score: 0.0882 | ...
   8. [Doc 88] ✅ ACERTO | Score: 0.0862 | ...
   9. [Doc 1008] ❌ FALLO | Score: 0.0857 | ...
   10. [Doc 270] ✅ ACERTO | Score: 0.0833 | ...
   RESUMEN: 6/10 relevantes encontrados.
```

```
>> MODELO 2: TF-IDF (Vectorial) (Top 10):
   1. [Doc 500] ✅ ACERTO | Score: 0.4689 | ...
   2. [Doc 450] ❌ FALLO | Score: 0.1880 | ...
   3. [Doc 87] ✅ ACERTO | Score: 0.1796 | ...
   4. [Doc 407] ✅ ACERTO | Score: 0.1741 | ...
   5. [Doc 408] ✅ ACERTO | Score: 0.1681 | ...
   6. [Doc 268] ✅ ACERTO | Score: 0.1658 | ...
   7. [Doc 88] ✅ ACERTO | Score: 0.1449 | ...
   8. [Doc 584] ❌ FALLO | Score: 0.1441 | ...
   9. [Doc 270] ✅ ACERTO | Score: 0.1361 | ...
   10. [Doc 607] ❌ FALLO | Score: 0.1351 | ...
   RESUMEN: 7/10 relevantes encontrados.
```

```
>> MODELO 3: BM25 (Probabilístico) (Top 10):
   1. [Doc 500] ✅ ACERTO | Score: 31.6903 | ...
   2. [Doc 268] ✅ ACERTO | Score: 16.6599 | ...
   3. [Doc 88] ✅ ACERTO | Score: 16.2980 | ...
   4. [Doc 270] ✅ ACERTO | Score: 16.1909 | ...
   5. [Doc 87] ✅ ACERTO | Score: 16.1127 | ...
   6. [Doc 407] ✅ ACERTO | Score: 15.7140 | ...
   7. [Doc 450] ❌ FALLO | Score: 15.7111 | ...
   8. [Doc 267] ✅ ACERTO | Score: 14.0176 | ...
   9. [Doc 408] ✅ ACERTO | Score: 13.0451 | ...
   10. [Doc 396] ❌ FALLO | Score: 13.0249 | ...
   RESUMEN: 8/10 relevantes encontrados.
```

Los resultados continuaron evidenciando el rendimiento superior de BM25 y a JACCARD como el menos acertivo.

5. Análisis de Métricas de Evaluación

La evaluación cuantitativa se realizó contrastando los rankings generados contra el *Ground Truth* (Qrels). Se utilizó una profundidad de recuperación de **Top-200** para asegurar un cálculo preciso del MAP, capturando documentos relevantes en la "larga cola".

Modelo	MAP (Calidad Global)	Precision@10	Recall@10
Jaccard	0.1767	0.2071	0.2016
TF-IDF	0.2548	0.2862	0.2721
BM25	**0.2706**	**0.2956**	**0.2799**

Interpretación de los Resultados:

1. **Superioridad del Modelo Probabilístico:** El modelo **BM25** obtuvo el mejor desempeño en todas las métricas (MAP: 0.27). Esto valida la hipótesis de diseño: la función de saturación no lineal de BM25 gestiona mejor la densidad léxica de los textos de ingeniería que el enfoque vectorial estándar.
2. **Exactitud en la Primera Página (P@10):** BM25 logró una precisión del **~30% en el Top 10**. Esto implica que, en promedio, 3 de cada 10 documentos mostrados en la primera página son estrictamente relevantes, un resultado competitivo considerando la especificidad del dominio.
3. **Insuficiencia del Enfoque Binario:** El modelo Jaccard (MAP 0.17) demostró ser insuficiente para este tipo de corpus. Al ignorar la frecuencia de los términos, no pudo distinguir entre un documento que menciona un concepto de pasada y uno que lo trata en profundidad.
4. **Impacto de la Limpieza de Stopwords:** La mejora observada respecto a pruebas iniciales (donde el MAP rondaba 0.16 en otros corpus) se atribuye directamente a la eliminación de las "Stopwords de Ingeniería". Al remover términos como *method* o *experimental*, se eliminaron falsos positivos que compartían metodología pero no temática.

Conclusión

El sistema desarrollado cumple con los requisitos de un motor de búsqueda académico funcional. La combinación de un preprocesamiento consciente del dominio y la calibración de un modelo probabilístico (BM25) permitió superar las líneas base clásicas, demostrando que la adaptación al tipo de vocabulario (Ingeniería) es tan crítica como la elección del algoritmo matemático.