ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE INGENIERÍA EN SISTEMAS
Departamento de Informática y
Ciencias de la Computación

RECUPERACIÓN DE LA INFORMACIÓN
Informe Proyecto 1er Bimestre
Sistema de Recuperación de Información

PERÍODO ACADÉMICO: 2025-A **GRUPO:** 8

PARALELO: GR1CC

PROFESOR: Ing. Iván Carrera

INTEGRANTES:

- Wilson Inga
- Anthony Reinoso
- Sergio Vite

FECHA DE ENTREGA: 09-06-2025

Objetivo.

1. Diseñar e implementar un sistema de recuperación de información que indexe un conjunto de documentos en texto plano.
2. Ejecutar consultas de texto libre utilizando el modelo vectorial con ponderación TF-IDF o BM25.
3. El sistema debe permitir evaluar la calidad de los resultados utilizando métricas estándar como precisión y recall.

Descripción del conjunto de datos: beir/arguana

- **Nombre:** BEIR - ArguAna
- **Tarea principal:** Recuperación de Argumentos
- **Descripción:** El conjunto de datos ArguAna consiste en preguntas argumentativas y argumentos relacionados obtenidos de sitios como *DebatePortal*. Dado un tema de debate (como "¿La pena de muerte debería ser abolida?"), el objetivo es recuperar los argumentos más relevantes y bien fundamentados desde una colección de textos argumentativos.
- **Tamaño:**
    - Aproximadamente 1,400 queries
    - Y una colección de alrededor de 8,000 documentos.
- **Idioma:** Inglés

Aplicación típica:

Este dataset es ideal para evaluar y entrenar sistemas de recuperación semántica, como BM25, DPR, TCT-ColBERT, o cualquier modelo capaz de capturar relaciones complejas entre argumentos y temas. Dado su enfoque argumentativo, se requiere una mayor comprensión del contenido semántico y de postura en los textos, lo que lo hace desafiante y útil para el desarrollo de modelos avanzados de búsqueda.

A continuación, se describen los tres componentes principales del conjunto de datos:

1. Consultas (Queries)

- Las consultas representan preguntas o temas argumentativos formulados en lenguaje natural. Estas consultas actúan como punto de partida para recuperar argumentos relevantes desde una colección de documentos.
- **Estructura**:

```
# ID único de la consulta
query.query_id = '1'
# Texto de la consulta
query.text  = 'Should animals be used for scientific experiments?'
```

2. Documentos (Documents)

- Los documentos corresponden a argumentos individuales extraídos de fuentes en línea dedicadas al debate, como DebatePortal. Cada documento contiene una postura o idea relacionada con temas argumentativos.
- **Estructura:**

```
{
# Identificador único del documento
 "doc_id": "D1234",
# Título del documento
  "title": null,
# Cuerpo textual del argumento.
  "text": "Animal testing has played a key role in the development of modern medicine. Without it, many life-saving treatments would not exist."
```

3. Juicios de Relevancia (Qrels)

- Los juicios de relevancia (qrels) indican la relación de relevancia entre cada consulta y los documentos. Estos juicios fueron generados manualmente para establecer una base de evaluación objetiva.

- **Estructura:**

```
{
 # Identificador de la consulta.
"query_id": "1",
# Identificador del documento evaluado.
  "doc_id": "D1234",
# Valor de relevancia (típicamente 1 = relevante).
  "relevance": 1
}
```

Explicación de las decisiones de diseño.

- Se utilizó Python como lenguaje principal con las bibliotecas Flask, Scikit-learn, Pandas y NLTK.
- El sistema realiza preprocesamiento sobre consultas y documentos: tokenización, lowercasing, stopwords, stemming y lematización.
- La evaluación se automatizó en base a archivos `qrels.tsv` y `queries_preprocessed.tsv`.
- La interfaz permite al usuario ingresar consultas libremente y seleccionar el modelo (TF-IDF o BM25).

Ejemplos de consultas y resultados.

Consulta: 'Animal research is only used when it's needed'

Consulta preprocesada: 'anim research use need eu'

Documento:
Animal research is only used when it's needed EU member states and the US have laws to stop animals being used for research if there is any alternative. The 3Rs principles are commonly used. Animal testing is being Refined for better results and less suffering, Replaced, and Reduced in terms of the number of animals used. This means that less animals have to suffer, and the research is better.

Consulta: 'That is, of course, not to say that children everywhere cannot be a cause for joy'

Consulta procesada: 'cours say child everywher caus joy'

Documento:

Any body of values that claims to respect the rights of the individual must recognise the right of a woman to choose Even the doctrines of the Church accepts that pregnancy is not, in and of itself, a virtue – there is no compulsion to maximise the number of pregnancies; there is simply

a disagreement about how they should be avoided. The Church recommends that couples may minimise the chance without ever making it impossible through a chemical or physical barrier. In some parts of the world a pregnancy, even one that is not planned, is seen as a time for joy – a blessing for the family that will lead to a new and happy life bringing pleasure to both parents, their society and the child. That ideal is very far from the experience of much of the world where a child is another mouth to feed on impossibly little income. For all too much of the world, that life will be cruel, nasty and short. In slums, favellas and barren wastes that life is likely to be one marked more by dysentery or diarrhea, malnutrition and misery than by the sanitised, idealised image promoted in the West. That is, of course, not to say that children everywhere cannot be a cause for joy, of course they can. Indeed even within the poorest of situations, a new child can be the focus of great joy in an otherwise hard life. However, if that is to be the case, that child must be planned and prepared for. Overwhelmingly, the mother is likely to have paramount responsibility for the child; so that planning and preparation needs to be theirs. It is difficult to imagine the scenario that would reach the objective observer to reach the conclusion that the right group of individuals to reach that decision were a group of celibate men who had never met the parents and would take to role in the care or support of the child. Yet that, astonishingly, is what Proposition would like us to believe.

Consulta: ‘A practice that is thousands of years old and has not been found to’

Consulta procesada: ‘practic thousand year old find’
Documento:

16 year olds are mature enough to vote 16 year olds are mature enough to make important decisions such as voting. If the government agrees that 16 year olds can have sex, join the army, and apply for a passport, then surely they are mature and responsible enough to decide who runs their country and makes important decisions that affect them. Their bodies are fully adult, they have been educated for at least 10 years, and most of them have some experience of work as well as school. By this time, it is likely a teenager will have developed “Advanced reasoning skills…the ability to think about multiple options and possibilities. It includes a more logical thought process and the ability to think about things hypothetically”. [1] This means they are able to form political views and they should be allowed to put these across at election time. Indeed by 16 children are as tolerant as adults and their political skill (the perceived ability to participate effectively in civil life by writing to political leaders and by speaking publically at meetings) is as high at 16 as for those in their late twenties. [2] There is no magic difference between 16 and 18 - indeed, many 16 year olds are more sensible than some 20 year olds. [1] Morgan, Erin, and Huebner, Angela, ‘Adolescent Growth and Development’, VirginiaTech, 1 Mary 2009 [2] Atkins, Robert, and Hart, Daniel, ‘American Sixteen and Seventeen Year Olds are Ready to Vote’, The ANNALS of the American Academy of Political and Social Science, Vol 633:201, 2011, p.210

Análisis de métricas de evaluación.

Para evaluar la calidad del sistema de recuperación de información, se utilizaron métricas estándar en el área de recuperación de texto: precisión (precision) y exhaustividad (recall).

Definiciones

- **Precision:** Mide la proporción de documentos relevantes entre los documentos recuperados.

$$precision = \frac{Documentos\ relevantes\ recuperados}{Total\ de\ documentos\ recuperados}$$

- **Recall:** Mide la proporción de documentos relevantes que fueron recuperados respecto al total de documentos relevantes existentes.

$$recall = \frac{documentos\ relevantes\ recuperados}{total\ de\ documentos\ relevantes\ en\ la\ coleccion}$$

Cálculo de métricas

Se realizaron pruebas utilizando un subconjunto de consultas del dataset **BEIR - ArguAna,** y se evaluaron los resultados en función de los juicios de relevancia (`qrels`) proporcionados por el conjunto de datos.

```
{
 # Identificador de la consulta.
"qrel.query_id": "1",
# Identificador del documento evaluado.
 "qrel.doc_id": "D1234",
# Valor de relevancia (típicamente 1 = relevante).
 "qrel.relevance": 1
}
```

Conclusiones

El análisis de las métricas confirma que el modelo BM25 es mas eficaz que el TF-IDF en la recuperación de argumentos relevantes en este tipo de corpus. Esto se debe, en parte a su capacidad de ajustar la ponderación de términos considerando la longitud del documento y la frecuencia global de los términos en el corpus.