Informe — Sistema de Recuperación de Información

Josune Singaña

10 de diciembre de 2025

Contents

1 Resumen

Este proyecto implementa un sistema de Recuperación de Información. El sistema permite indexar un corpus, ejecutar consultas de texto libre y evaluar

los resultados mediante métricas estándar.

El corpus utilizado es **Amazon Fine Food Reviews**, empleando únicamente la columna `Text` (contenido de la reseña)

2 Corpus

El corpus contiene miles de reseñas escritas por usuarios. Solo se utiliza el texto principal de cada reseña. El usuario puede limpiar manualmente el CSV antes de usarlo.

El módulo `index.py` permite cargar el corpus desde CSV y obtener una lista de documentos.

3 Preprocesamiento

Se implementa un **preprocesamiento fuerte** en `index.py` mediante la función:

- Conversión a minúsculas
- Eliminación de HTML
- Eliminación de caracteres especiales
- Normalización de espacios
- Tokenización por espacios
- Stopwords básicas

El objetivo es obtener tokens representativos y homogéneos.

4 Diseño

5 Construcción del índice

El archivo `index.py` incluye:

- Tokenizador
- Diccionario invertido: término $\rightarrow$ {doc: frecuencia}
- Cálculo de longitudes documentales

Este índice sirve como base para los modelos Jaccard, TF-IDF y BM25.

6 Modelos de Recuperación

6.1 Jaccard (binario)

Mide intersección sobre unión entre tokens de consulta y documento.

6.2 TF-IDF (vectorial)

Se utiliza `TfidfVectorizer` de scikit-learn por eficiencia y robustez. Las similitudes se calculan con coseno.

6.3 BM25

Implementación propia:

- Parámetros `k1=1.5` y `b=0.75`
- Cálculo de IDF con suavizado
- Normalización por longitud del documento

7 Interfaces

7.1 CLI (archivo cli.py)

Permite ejecutar:

```
python cli.py --corpus amazon.csv --model bm25 --query "great taste"
```

7.2 Interfaz simple (archivo $\text{interface}_{\text{simple.py}}$)

Muestra un menú interactivo que facilita la ejecución de consultas.

8 Evaluación

El módulo `evaluation.py` permite calcular:

- Precision@k
- Recall@k
- Average Precision (AP)
- Mean Average Precision (MAP)

El archivo `qrels_utils.py` permite:

- Cargar qrels reales
- Generar qrels de ejemplo

9 Conclusión

Este sistema cumple con todas las especificaciones solicitadas:

- Indexado
- Preprocesamiento
- Modelos Jaccard, TF-IDF y BM25
- Interfaz CLI y menú simple
- Evaluación con qrels