

Proyecto de Recuperación de Información

Sistemas de Recuperación de Información

Curso: 2024-2025

Semestre: 2do

El objetivo principal del proyecto consiste en diseñar e implementar, de manera individual, un modelo funcional de recuperación de información, partiendo de los fundamentos teóricos estudiados y aplicando al menos una mejora sobre el enfoque seleccionado.

A continuación se describen por secciones aspectos importantes del proyecto.

Información general

Cuando el estudiante decida comenzar a trabajar en la solución del proyecto, deberá seguir los siguientes pasos:

1. Leer detenidamente este documento en su totalidad.
2. Investigar a profundidad el modelo de recuperación de información seleccionado.
3. Implementar todas las funciones especificadas en el repositorio.
4. Completar la información requerida.
5. Presentar y defender el trabajo ante los evaluadores, si es requerido.

La calificación final se determinará con base en los siguientes componentes:

- Calidad técnica de la implementación del modelo.
- Posición en el ranking según las métricas establecidas.
- Completitud y claridad de la documentación.
- Solidez en la discusión y defensa del trabajo realizado (cuando aplique).

Modelos de Recuperación de Información

Los estudiantes podrán seleccionar con el objetivo de implementar cualquier modelo no abordado en clase, o sea, aquellos que representen extensiones o variaciones de los modelos clásicos de recuperación de información. Las opciones válidas incluyen:

- Modelo Booleano Difuso (Fuzzy Boolean)
- Modelo Booleano Extendido (Extended Boolean)
- Modelo Vectorial Generalizado (Generalized Vector Space)
- Modelo de Semántica Latente (Latent Semantic Indexing)
- Modelo basado en Redes Neuronales (Neural IR Models)
- Modelo de Redes de Inferencia (Inference Networks)
- Modelo de Redes de Creencia (Belief Networks)

Importante: La selección del modelo deberá:

- Incluir referencia a fuentes académicas válidas (artículos, libros o publicaciones reconocidas).
- Mantener coherencia con los principios teóricos de recuperación de información.

Instrucciones para la implementación

Siga cuidadosamente los siguientes pasos para comenzar su implementación:

1. **Clonar el repositorio:** Realice una copia local del proyecto desde <https://github.com/SRI-MatCom/ir-core->

2. **Configurar archivo personal**

- Duplique el archivo `template.py`.
- Renómbrelo siguiendo la estructura: `Nombre_Apellido1_Apellido2-Carrera-Grupo.py` donde:
 - **Carrera:** D (Ciencia de Datos) o C (Ciencia de la Computación).
 - **Grupo:** 311 o 312 según corresponda.

Ejemplo: Para el estudiante Francisco Guerra Piloto del grupo 311 de Computación, su fichero quedaría `Francisco_Guerra_Piloto_C_311.py`

3. **Personalizar el archivo**

- Complete todas las variables globales con sus datos personales, la cuales incluyen:
 - **NAME:** indica el nombre completo del estudiante.
 - **GROUP:** indica el grupo al que pertenece el estudiante.
 - **CAREER:** indica la carrera del estudiante.
 - **MODEL:** indica el nombre del modelo implementado.

4. **Implementar funcionalidad**

- Modifique, según el modelo, las funciones `fit` y `predict`.
La función `fit` está creada para que se encargue de las transformaciones necesarias del corpus. Por su parte, `predict` determina los documentos más relevantes por cada consulta del corpus especificado.
 - Puede (y se recomienda) crear funciones auxiliares prefijadas con `_` (ej: `_preprocesar_texto`) para mantener un código modular y desacoplado, siguiendo las convenciones de los métodos “privados”.
 - Implemente manejo robusto de excepciones.

Restricciones programáticas

El proyecto permite total libertad de implementación, con la única condición de utilizar exclusivamente las bibliotecas de Python especificadas en el archivo `requirements.txt`. No se aceptarán dependencias adicionales.

Posibles mejoras

Como desarrollador, tiene plena autonomía para implementar cualquier mejora que tribute al rendimiento de su modelo, siempre que:

- **No altere** la definición fundamental del modelo seleccionado.
- **Mantenga** la interfaz de las funciones especificadas.
- **Documente** claramente todas las modificaciones realizadas.

Algunas mejoras a tener en cuenta, inspiradas en las conferencias de la asignatura pueden ser:

- Técnicas avanzadas de preprocesamiento de texto.
- Mecanismos de ponderación o normalización.

- Estrategias de expansión de consultas.
- Retroalimentación del modelo.

La implementación de mejoras será valorada positivamente en la evaluación, especialmente cuando:

- Demuestre comprensión teórica.
- Justifique técnicamente su elección.
- Muestre impacto medible en los resultados.

Sobre el informe

Dado que no se exigirá un informe separado, deberá incluir en su archivo (entre los datos personales y la definición de clase) un *string* con la siguiente información estructurada:

- **Fuente bibliográfica principal:** Referencia completa (autor, título, año, url) que fundamenta su implementación.
- **Mejoras implementadas** (si aplica):
 - Descripción técnica concisa (máx. 3 líneas).
 - Beneficio esperado en el modelo.
- **Análisis conceptual** (respuestas breves):
 - **Definición formal:** Caracterización matemática del modelo.
 - **Dependencias entre términos:** Sí/No + explicación (1 línea).
 - **Correspondencia parcial documento-consulta:** Sí/No + explicación (1 línea).
 - **Ranking:** Sí/No + explicación (1 línea).

Entrega

El plazo tope para la entrega del fichero implementado es el **28 de abril** a las **4:00pm**.

Cada estudiante debe de poner en el grupo de telegram <https://t.me/+WbYyRpqsbxAwNzMx> su solución.