# Recuperación de la información

Curso 2024/2025
Grado en Ingeniería Informática – Ingeniería del Software
Propuesta de trabajo práctico de la asignatura Inteligencia Artificial

Prof. Álvaro Romero Jiménez

## Introducción y objetivos

Como campo académico de estudio, la recuperación de información se puede definir como el proceso de buscar y encontrar material (normalmente documentos) de naturaleza no estructurada (normalmente texto) que satisfaga una necesidad de información, donde la búsqueda se realiza dentro de grandes colecciones (normalmente almacenadas en ordenadores) de documentos.

Definida de este modo, la recuperación de información solía ser una actividad a la que solo se dedicaban unas pocas personas: bibliotecarios referencistas, asistentes jurídicos y buscadores profesionales similares. En la actualidad, el mundo ha cambiado y cientos de millones de personas se involucran cada día en la recuperación de información cuando utilizan un motor de búsqueda web o consultan su correo electrónico. La recuperación de información se está convirtiendo rápidamente en la forma dominante de acceso a la información, superando a la búsqueda tradicional en bases de datos.

El **objetivo principal** de esta propuesta es la construcción y evaluación de varios sistemas de recuperación de la información que permitan satisfacer necesidades de información sobre un conjunto determinado de documentos.

Para ello será necesario alcanzar los siguientes **objetivos específicos**:

1. Recopilar un corpus de documentos concreto para el trabajo.
2. Diseñar una colección de necesidades de información para usar en la evaluación de los sistemas de recuperación de información que se construyan.
3. Construir y evaluar al menos un sistema de recuperación de información que admita consultas especificadas según el modelo booleano, seleccionando los documentos del corpus compatibles con ellas.
4. Construir y evaluar al menos un sistema de recuperación de información que admita consultas especificadas en texto libre, estableciendo en relación a ellas un *ranking* entre los documentos del corpus.
5. Documentar el trabajo realizado usando un formato de artículo científico.
6. Realizar una presentación (PDF, PowerPoint o similar) de los resultados obtenidos.

# Descripción del trabajo

A continuación se describe con más detalle cómo debe llevarse a cabo el trabajo.

La referencia bibliográfica principal de la propuesta es el libro Introduction to Information Retrieval (IIR), que es al que hacen alusión los capítulos y secciones indicadas en lo que sigue.

## Metodología

En primer lugar debe recopilarse un corpus de documentos de texto (IIR, sección 2.1) sobre una determinada temática (recetas de cocinas, poemas de un determinado autor, fichas de jugadores de fútbol o cualquier otra temática de interés para los alumnos). **El corpus debe ser validado por el profesor**, lo que se recomienda hacer antes de realizar la tarea de recopilación de documentos, y debe contener obligatoriamente al menos cincuenta documentos.

A continuación debe diseñarse una colección de necesidades de información (por ejemplo, una necesidad de información podría ser «recetas aptas para personas celíacas») que se usará para la evaluación de los sistemas de recuperación de la información que se construyan. Esta colección deberá contener obligatoriamente al menos veinte necesidades de información y para cada una de ellas deben identificarse cuáles de los documentos del corpus son relevantes, es decir, contienen información que un usuario consideraría relacionada con su necesidad de información (por ejemplo, las recetas sin gluten) y cuáles son no relevantes (por ejemplo, las recetas con gluten).

Respecto a los sistemas de recuperación de la información que se piden en este trabajo, la primera tarea a realizar es la construcción de un índice invertido (IIR, sección 1.2) que relacione cada término del vocabulario con los documentos que lo contienen. Debe considerarse también la conveniencia de incluir en el índice información adicional que facilite y haga más eficiente las posteriores consultas al sistema, así como el uso de técnicas de normalización de términos como la eliminación de palabras vacías o la aplicación de procedimientos de reducción a la raíz y de lematización (IIR, sección 2.2).

Finalmente, deben construirse

- Un sistema de recuperación de la información (o varios, si por ejemplo se pretende evaluar el efecto de las distintas técnicas de normalización de términos) que admita consultas especificadas según el modelo booleano, es decir, combinando términos individuales con los operadores lógicos AND, OR y NOT, y devuelva todos los documentos del corpus compatibles con la consulta (IIR, capítulo 1). Los sistemas construidos se evaluarán mediante la sensibilidad y la precisión (IIR, sección 8.3) promediadas sobre las necesidades de información diseñadas anteriormente, para lo que para cada una de ellas deberán derivarse consultas booleanas adecuadas.
- Uno o varios sistemas de recuperación de la información que admitan consultas especificadas en texto libre y usen el modelo tf-idf y la similitud del coseno entre la consulta y los documentos del corpus para devolver estos en el orden determinado por la mencionada similitud (IIR, secciones 1.4, 6.2 y 6.3). Los sistemas construidos se evaluarán mediante MAP (IIR, sección 8.4), la precisión media promediada sobre las necesidades de información diseñadas anteriormente, para lo que para cada una de ellas deberán derivarse consultas de texto libre adecuadas.

Para cada uno de los sistemas de recuperación de la información construidos debe implementarse una interfaz (no es necesario que sea gráfica, basta con que sea textual) para la realización de consultas sobre el corpus.

**En cada convocatoria debe recopilarse un corpus distinto** y, en consecuencia, debe diseñarse una colección distinta de necesidades de información.

## Lenguaje de programación y bibliotecas recomendadas

Para la realización del trabajo debe utilizarse el lenguaje de programación Python.

Es posible implementar desde cero todos los elementos de los sistemas de recuperación de información que se piden en esta propuesta, pero es más conveniente usar bibliotecas de Python ya disponibles, entre las que se recomiendan la siguientes:

- Beautiful Soup, facilita la extracción de texto contenido en documentos HTML.
- NLTK, proporciona herramientas para el análisis y la normalización de texto.
- Whoosh, facilita la creación de índices invertidos y la realización de consultas de búsqueda.

## Documentación y entrega

El trabajo deberá documentarse siguiendo un formato de artículo científico, con una **extensión mínima de 6 páginas**. En la página web de la asignatura se pueden encontrar plantillas donde se sugiere una estructura general. Estas plantillas siguen el formato de los *IEEE conference proceedings*, en cuyo sitio web la guía para autores ofrece información más detallada. El documento entregado deberá estar en formato PDF. Se valorará el uso del sistema LaTeX.

En el caso concreto de este trabajo, la memoria deberá al menos incluir: introducción; descripción del corpus recopilado y de la colección de necesidades de información diseñada; descripción de los sistemas de recuperación de información construidos, explicando las dificultades encontradas y las decisiones de diseño adoptadas para abordarlas; descripción de los resultados alcanzados; conclusiones; bibliografía. **En ningún caso debe incluirse código en la memoria**.

La entrega del trabajo consistirá de **un único fichero comprimido `zip`** conteniendo la memoria del trabajo, el corpus de documentos, la colección de necesidades de información diseñada, junto con las consultas derivadas a partir de ellas para la evaluación de los sistemas y el código implementado (ficheros py o cuadernos de Jupyter). Si el corpus de documentos resultara demasiado grande para poder subirlo a través de la página de la asignatura, entonces deberá subirse a Google Drive o similar, proporcionándose un enlace de compartición al mismo.

## Presentación y defensa

Como parte de la evaluación del trabajo se deberá realizar una defensa del mismo, para lo que se citará a los alumnos de manera conveniente.

Al inicio de la defensa se deberá realizar una pequeña presentación (PDF, PowerPoint o similar) de diez minutos en la que participarán activamente todos los miembros del grupo que ha desarrollado el trabajo. Esta presentación deberá seguir a grandes

rasgos la misma estructura que la memoria del trabajo, haciendo especial mención a los resultados obtenidos y al análisis crítico de los mismos.

En los siguientes diez minutos de la defensa, el profesor procederá a realizar preguntas sobre el trabajo, que podrán ser tanto de la memoria como del código fuente.

### Uso de inteligencia artificial generativa

El uso de sistemas de inteligencia artificial generativa está permitido con las siguientes condiciones:

- Debe explicarse para qué se han utilizado esos sistemas, así como describir las entradas (*prompts*) proporcionadas a los mismos.
- En la defensa del trabajo **ambos alumnos** deben demostrar conocimiento y entendimiento de la **totalidad del trabajo**, lo que incluye las respuestas obtenidas de esos sistemas.

## Evaluación del trabajo

Para la evaluación del trabajo se tendrán en cuenta los siguientes criterios, considerando una nota total máxima de 4 puntos:

- *Memoria del trabajo* (hasta 1 punto): se valorará la claridad de las explicaciones, el razonamiento de las decisiones, el análisis y presentación de resultados y el correcto uso del lenguaje. La elaboración de la memoria debe ser original, por lo que no se evaluará el trabajo si se detecta cualquier copia del contenido.
- *Corpus recopilado y colección de necesidades de información* (hasta 0.75 puntos): se valorará el tamaño y la adecuación de ambos.
- *Sistemas de recuperación de la información* (hasta 1.5 puntos): se valorará la implementación realizada, el uso de técnicas para tratar de mejorar el rendimiento de los sistemas y la facilidad de uso de las interfaces proporcionadas.
- *Código fuente* (hasta 0.75 puntos): se valorará la claridad y buen estilo de programación, corrección y eficiencia de la implementación y calidad de los comentarios. El código debe ser original, por lo que no se evaluará el trabajo si se detecta código copiado o descargado de internet.
- *Presentación y defensa*: se valorará la claridad de la presentación y la buena explicación de los contenidos del trabajo así como, especialmente, las respuestas a las preguntas realizadas por el profesor, lo que dará lugar para cada alumno por separado a un factor multiplicativo en el intervalo $[0, 1]$ de la nota total obtenida a partir de los apartados anteriores.

**IMPORTANTE**: cualquier plagio, compartición de código o uso de material que no sea original y del que no se cite convenientemente la fuente, significará automáticamente la calificación de cero en la asignatura para todos los alumnos involucrados. Por tanto, a estos alumnos no se les conserva, ni para la actual ni para futuras convocatorias, ninguna nota que hubiesen obtenido hasta el momento. Todo ello sin perjuicio de las correspondientes medidas disciplinarias que se pudieran tomar.