MEMORIA DEL PROYECTO: SISTEMA DE RECUPERACIÓN DE INFORMACIÓN

1. Introducción

Este proyecto desarrolla un **Sistema de Recuperación de Información** basado en el **Modelo de Espacio Vectorial**. Su objetivo es indexar una colección de documentos XML extraídos de SciELO y permitir consultas de búsqueda eficientes. El sistema procesa los documentos, los representa en un espacio vectorial y utiliza un modelo de similitud para recuperar los documentos más relevantes para una consulta dada.

Cada fase se implementa en una práctica independiente, integrándose en el **archivo principal (main.py)**.

---

2. Arquitectura del Sistema

El sistema está compuesto por varios módulos Python, cada uno implementando una funcionalidad específica. A continuación, se describen los archivos y sus responsabilidades:

2.1. Módulo Practica1_1.py: Carga y Preprocesamiento de Documentos

Este módulo se encarga de leer los archivos XML y extraer su contenido relevante.

- **Funciones clave:**
    - charge_config(config_file): Carga la configuración y obtiene la ruta de los documentos XML.
    - extract(file_path): Extrae los metadatos del documento: título, autores, fecha, palabras clave y texto completo.
    - transform(text): Normaliza y tokeniza el texto, convirtiéndolo en una lista de palabras.
    - load(file_name, tokens, output_dir): Guarda los tokens procesados en formato JSON.

2.2. Módulo Practica1_2.py: Eliminación de Stopwords

Este módulo filtra palabras vacías (stopwords) del texto procesado.

- **Función clave:**
    - stopper(tokens): Elimina las stopwords en español utilizando la biblioteca nltk.

2.3. Módulo Practica1_3.py: Stemming

Aquí se aplica el algoritmo de **SnowballStemmer** para reducir las palabras a su raíz léxica.

- **Función clave:**
    - stem_words(palabras): Devuelve la lista de palabras convertidas a sus raíces.

2.4. Módulo Practica1_4.py: Diccionarios e Índice Invertido

Este módulo asigna identificadores únicos a términos y documentos, y construye el índice invertido.

- **Funciones clave:**
    - enumeracion(words): Genera term2id e id2term, mapeando términos a identificadores.
    - enum_docs(files): Genera doc2id e id2doc, asignando IDs a los documentos.
    - indice_invertido(id2term, word, doc2id, file, tok_file): Construye el índice invertido registrando en qué documentos aparece cada término y con qué frecuencia.

2.5. Módulo Practica1_5.py: Cálculo de Pesos TF-IDF y Representación Vectorial

Aquí se calculan los pesos de los términos en los documentos utilizando el modelo **TF-IDF** y se construye la matriz de documentos.

- **Funciones clave:**
    - calcular_pesos(inverted_index, doc2id): Calcula el índice invertido con pesos normalizados y el IDF de cada término.
    - calcular_todo(inverted_index, doc2id, term2id): Ejecuta el flujo completo de cálculo de pesos, normalización y construcción de la matriz de documentos.

2.6. Módulo Practica1_6.py: Procesamiento de Consultas y Cálculo de Similitud

Este módulo permite realizar consultas sobre la colección y recuperar los documentos más relevantes mediante el **producto escalar**.

- **Funciones clave:**
    - procesar_consulta(text): Preprocesa la consulta (tokenización, eliminación de stopwords y stemming).
    - vectorizar_consulta(peso_consulta, term2id, mapping_tokens): Representa la consulta en el mismo espacio vectorial que los documentos.
    - dot_product(vector1, vector2): Calcula la similitud entre la consulta y cada documento.
    - buscar_consulta(query_text, max_docs, document_matrix, idf, term2id, id2doc): Devuelve los documentos más relevantes para la consulta.
    - procesar_consultas_desde_fichero(query_filename, max_docs, document_matrix, idf, term2id, id2doc): Procesa un conjunto de consultas desde un fichero.

2.7. Módulo Practica1_7.py: Presentación de Resultados

Este módulo muestra los resultados de las consultas en **dos formatos**:

1. **Formato amplio**: Muestra la consulta completa, el ranking, la similitud y el nombre del documento.
2. **Formato compacto**: Muestra solo el ID de la consulta y el ID del documento.

- **Funciones clave:**
    - resultados_amplios(query_results, queries): Presenta los resultados con información detallada.
    - resultados_compactos(query_results): Presenta los resultados en formato conciso.
    - retrieve_name(file, config_file): Recupera el nombre real de un documento desde su ID.

---

3. Flujo de Ejecución del Programa

El sistema se ejecuta a través de main.py, el cual orquesta todo el proceso:

1. **Carga y preprocesamiento de los documentos.**
2. **Creación de diccionarios de términos y documentos.**
3. **Construcción del índice invertido.**
4. **Cálculo de pesos TF-IDF y representación en una matriz vectorial.**
5. **Procesamiento de consultas y cálculo de similitud.**
6. **Visualización de resultados en distintos formatos.**

4. Estructuras de datos utilizadas

Las estructuras de datos que se han utilizado han sido sobre todo diccionarios y listas.

4.1 Listas utilizadas:

1. files: Guarda todos los nombres de los archivos de la colección de SciELO, la usamos para iterar sobre ella y conseguir acceder a cada archivo y extraer su información.

2. tokens: La primera lista con todos los tokens del archivo.

3. tokens1: La segunda lista con los tokens sin las stopwords.

4.tokens3

5. stem_tokens: donde están todos los tokens tras pasar el filtro del stemmer, con repetidos

6. indice_invertido: es una lista de diccionarios con las apariciones de la palabra en cada archivo.

7. queries: Son las consultas en raw.

4.2 Sets utilizados:

1. all_tokens: Este set almacena todos los tokens sin repetidos para más tarde usarlo para la creación de diccionarios.

4.3 Diccionarios utilizados:

1. term2id: Almacena como clave el token y guarda el id del token.

2. id2term: Almacena como clave el id y guarda el token.

3. doc2id: Almacena como clave el nombre de los archivo para acceder al id del mismo

4. id2doc: Almacena como clave el id del archivo para guardar el nombre del archivo.

5. indice_full: Es la estructura del índice invertido que relaciona cada palabra con sus documentos y cuantas veces aparece en cada uno de ellos, es una estructura enorme.

6. norm_index: Representa los pesos de cada palabra en cada archivo, sigue el mismo formato que indice_full, no es una “matriz” dispersa.

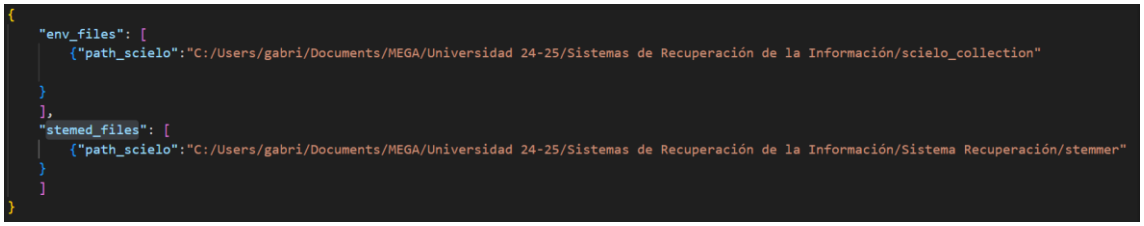
7. term_idf: Asocia cada id del token a su IDF calculado.

8. document_matrix: Cada clave es el ID de un documento y el valor es un diccionario que contiene, para cada término (según term2id), el peso normalizado si aparece o 0 en caso contrario.

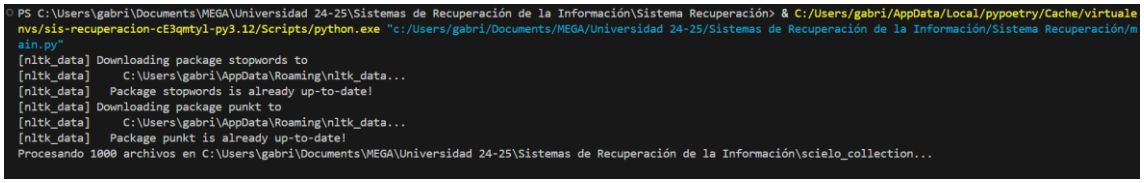
9. query_results: Representa las 10 mejores queries de cada consulta.

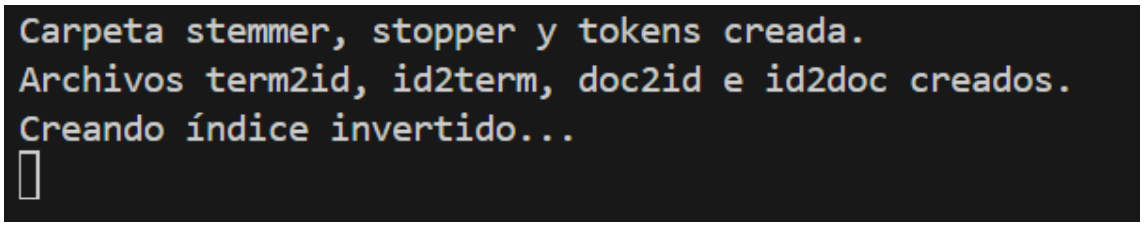
Los parámetros necesarios para la aplicación:

Es una estructura en .json llamado “config.json” que especifica la carpeta de los archivos de la colección con “env_files” y “stemed_files” carpeta donde se encuentran los archivos tras el stemmer.



Ejecución:





```
Índice invertido creado en 98.51 segundos.
Calculando pesos normalizados y generando matriz de documentos...
```

Y ahora el resto del texto:
La diabetes en personas mayores.

Similitud obtenida: S0212-97282014000100033.json - 0.41333370466673913

Nombre del documento: ¿Quién hizo qué?: diferencias entre adultos jóvenes y mayores en la memoria para un atraco

Similitud obtenida: S1699-695X2009000100004.json - 0.31732195309414923

Nombre del documento: Aproximación a la Diabetes Mellitus Oculta en un Servicio de Urgencias Hospitalario

Similitud obtenida: S0211-69952016000600535.json - 0.2834297251739145

Nombre del documento: Factores predictivos de nefropatía no diabética en pacientes diabéticos. Utilidad de la biopsia renal

Similitud obtenida: S0211-69952012000600012.json - 0.25950148095944353

Nombre del documento: Factores pronósticos de enfermedad coronaria en diabéticos asintomáticos para inclusión en lista de trasplante renal: Despistaje con coronariografía

Similitud obtenida: S1699-695X2011000100006.json - 0.23605589950265896

Nombre del documento: Control Metabólico en Pacientes Diabéticos Tipo 2: grado de Control y nivel de Conocimientos (Estudio AZUER)

Similitud obtenida: S0212-97282015000300027.json - 0.22815924731057602

Nombre del documento: Afrontamiento de problemas de salud en personas muy mayores

Similitud obtenida: S1699-695X2015000300004.json - 0.22396491684182238

Nombre del documento: Talleres sobre problemas de salud prevalentes con personas integradas en grupos de trabajo en centros de mayores

Similitud obtenida: S0211-69952011000200010.json - 0.21579895145372968

Nombre del documento: Diálisis peritoneal actual comparada con hemodiálisis: análisis de supervivencia a medio plazo en pacientes incidentes en diálisis en la Comunidad Canaria en los últimos años

Similitud obtenida: S0212-97282016000300023.json - 0.2103410965102888

Nombre del documento: Relación entre autoconcepto y nivel de depresión en personas con retinosis pigmentaria

Similitud obtenida: S1699-695X2017000100003.json - 0.2040701175413244

Nombre del documento: Evaluación de la prescripción de metformina en pacientes diabéticos tipo 2 de una institución de Atención Primaria en Salud en Cartagena de Indias, Colombia

Los jóvenes universitarios.

Similitud obtenida: S0212-97282013000100023.json - 0.3673653735603669

Nombre del documento: Ajuste social y escolar de jóvenes víctimas de maltrato infantil en situación de acogimiento residencial

Similitud obtenida: S0212-97282014000100033.json - 0.24406991094945926

Nombre del documento: ¿Quién hizo qué?: diferencias entre adultos jóvenes y mayores en la memoria para un atraco

Similitud obtenida: S1139-76322011000200005.json - 0.23786855720748262

Nombre del documento: Internet, sexo y adolescentes: una nueva realidad: Encuesta a jóvenes universitarios españoles

Similitud obtenida: S1699-695X2014000100003.json - 0.23185200419139293

Nombre del documento: Síntomas de depresión y ansiedad en jóvenes universitarios: prevalencia y factores relacionados

Similitud obtenida: S0212-97282013000100009.json - 0.21351588498977583

Nombre del documento: Autoeficacia en la prevención sexual del Sida: la influencia del género

Similitud obtenida: S1699-695X2010000300002.json - 0.20886003699460398

Nombre del documento: Hábitos, Preferencias y Satisfacción Sexual en Estudiantes Universitarios

Similitud obtenida: S1699-695X2009000200005.json - 0.20436308394841918

Nombre del documento: El Arte de Curar: estudio sobre Vías de Administración. Diferencias entre medio rural y urbano

Similitud obtenida: S0212-97282013000100021.json - 0.19352357260564018

Nombre del documento: Prevalencia de acontecimientos potencialmente traumáticos en universitarios españoles

Similitud obtenida: S1699-695X2010000300004.json - 0.18847046982708504

Nombre del documento: Historia de Embarazos en Estudiantes de Programas de Salud en una Universidad Pública del Caribe Colombiano

Similitud obtenida: S0212-97282014000300029.json - 0.18842723826444943

Nombre del documento: Búsqueda de sensaciones y consumo de alcohol: el papel mediador de la percepción de riesgos y beneficios

La insuficiencia renal.

Similitud obtenida: S0211-69952015000400007.json - 0.4728803985215492

Nombre del documento: Repetición de la medición de creatinina sérica en atención primaria: no todos tienen insuficiencia renal crónica

Similitud obtenida: S0211-69952010000100011.json - 0.3968363547989203

Nombre del documento: Análisis clínico de una población con poliquistosis renal autosómica dominante

Similitud obtenida: S0211-69952012000200011.json - 0.3511392472812198

Nombre del documento: Progresión de la enfermedad renal crónica en pacientes con enfermedad poliquística autosómica dominante

Similitud obtenida: S0211-69952014000100014.json - 0.2558806607353482

Nombre del documento: Microhematuria persistente con proteinuria negativa o de escasa cuantía

Similitud obtenida: S0211-69952012000100009.json - 0.25032104603497135

Nombre del documento: El cálculo de la creatinina sérica basal sobrestima el diagnóstico de alteración renal aguda en pacientes operados de cirugía cardíaca

Similitud obtenida: S0211-69952016000700609.json - 0.23222872711734166

Nombre del documento: La fragilidad en el anciano con enfermedad renal crónica

Similitud obtenida: S0211-69952009000600013.json - 0.22768463663742106

Nombre del documento: Análisis genético (PKD2) de la poliquistosis renal autosómica dominante

Similitud obtenida: S0211-69952013000500009.json - 0.2224228594272918

Nombre del documento: Tratamiento con hemodiálisis larga con filtros de alto cut-off en la nefropatía por cilindros del mieloma: nuestra experiencia

Similitud obtenida: S0211-69952012000500007.json - 0.22215303893541943

Nombre del documento: Enfermedad arterial periférica e insuficiencia renal: una asociación frecuente

Similitud obtenida: S0211-69952010000300008.json - 0.220326772923726

Nombre del documento: Enfermedad renal ateroembólica: un análisis de los factores clínicos y terapéuticos que influyen en su evolución

El asma un importante factor para predecir virus.

Similitud obtenida: S1139-76322010000500010.json - 0.6269913532605416

Nombre del documento: Claves de educación en asma: casos clínicos interactivos

Similitud obtenida: S1699-695X2013000300004.json - 0.48134726624713203

Nombre del documento: Efectividad de un programa de terapia de familia en niños asmáticos con familias disfuncionales

Similitud obtenida: S1139-76322009000100007.json - 0.4407484257066273

Nombre del documento: Tratamiento de las sibilancias recurrentes: asma en el niño menor de 3 años de edad

Similitud obtenida: S1139-76322014000400013.json - 0.33597762944801324

Nombre del documento: La gripe y las vacunas frente a la gripe: presente y futuro

Similitud obtenida: S1139-76322013000200003.json - 0.32808378907531

Nombre del documento: Evaluación de los conocimientos paternos sobre asma con el Newcastle Asthma Knowledge Questionnaire

Similitud obtenida: S0212-97282015000200009.json - 0.2758444093846843

Nombre del documento: Positividad y afrontamiento en pacientes con trastorno adaptativo

Similitud obtenida: S0212-97282014000200037.json - 0.24153038022381265

Nombre del documento: Teoría de la Acción Planeada y tasa de ejercicio percibida: un modelo predictivo en estudiantes adolescentes de educación física

Similitud obtenida: S1139-76322009000400002.json - 0.206816939124439

Nombre del documento: Nueva gripe [A(H1N1) 2009]: definición de caso sospechoso. Revisión de la concordancia en los criterios de definición de caso utilizados en las distintas comunidades autónomas españolas

Similitud obtenida: S0211-69952016000600510.json - 0.19913287263952562

Nombre del documento: Importancia relativa de los factores determinantes de los niveles séricos de 25-hidroxi-colecalciferol en la enfermedad renal crónica

Similitud obtenida: S1139-76322009000200005.json - 0.18401602840778164

Nombre del documento: Programa del asma en Atención Primaria: estudio comparativo entre dos centros de salud de Valladolid

Algunas enfermedades respiratorias, infecciosas e intestinales.

Similitud obtenida: S1699-695X2011000100005.json - 0.6922244380197982

Nombre del documento: Estudio epidemiológico en el Área de Salud de Entre Ríos

Similitud obtenida: S0211-69952017000100009.json - 0.5712624696670899

Nombre del documento: Microbiota intestinal en la enfermedad renal crónica

Similitud obtenida: S1139-76322014000400014.json - 0.5164982428795404

Nombre del documento: Biomarcadores para el despistaje de enfermedades infecciosas: una revolución diagnóstica para los países pobres

Similitud obtenida: S1699-695X2009000300004.json - 0.34774131404216496

Nombre del documento: Contaminación Atmosférica, Morbilidad y Mortalidad en la ciudad de Albacete (Año 2005)

Similitud obtenida: S1139-76322010000500002.json - 0.3307034569408254

Nombre del documento: Recomendaciones de la Conferencia de Consenso de Bronquiolitis Aguda en España: de la evidencia a la práctica

Similitud obtenida: S0211-69952015000100002.json - 0.2631071751648632

Nombre del documento: ¿Cuándo debe sospechar un nefrólogo una enfermedad mitocondrial?

Similitud obtenida: S1135-76062002000200007.json - 0.25271560793744274

Nombre del documento: La causa y la manera de la muerte indeterminada: a propósito de un caso de muerte súbita en adolescente, portador de una tumoración quística intestinal, descubierta durante la autopsia

Similitud obtenida: S1139-76322012000500004.json - 0.21057005523148348

Nombre del documento: Influencia de la asistencia a guarderías sobre la morbilidad en niños menores de 12 meses de edad

Similitud obtenida: S1139-76322016000400007.json - 0.18936647716524518

Nombre del documento: Adecuación del diagnóstico y tratamiento de la faringoamigdalitis aguda a las guías actuales

Similitud obtenida: S0212-97282016000100003.json - 0.18447901140714174

Nombre del documento: Cómo vivir con EPOC: percepción de los pacientes

1 S0212-97282014000100033.json

1 S1699-695X2009000100004.json

1 S0211-69952016000600535.json

1 S0211-69952012000600012.json

1 S1699-695X2011000100006.json

1 S0212-97282015000300027.json

1 S1699-695X2015000300004.json

1 S0211-69952011000200010.json

1 S0212-97282016000300023.json

1 S1699-695X2017000100003.json

2 S0212-97282013000100023.json

2 S0212-97282014000100033.json

2 S1139-76322011000200005.json

2 S1699-695X2014000100003.json

2 S0212-97282013000100009.json

2 S1699-695X2010000300002.json

2 S1699-695X2009000200005.json

2 S0212-97282013000100021.json

2 S1699-695X2010000300004.json

2 S0212-97282014000300029.json

3 S0211-69952015000400007.json

3 S0211-69952010000100011.json

3 S0211-69952012000200011.json

3 S0211-69952014000100014.json

3 S0211-69952012000100009.json

3 S0211-69952016000700609.json

3 S0211-69952009000600013.json

3 S0211-69952013000500009.json

3 S0211-69952012000500007.json

3 S0211-69952010000300008.json

4 S1139-76322010000500010.json

4 S1699-695X2013000300004.json

4 S1139-76322009000100007.json

4 S1139-76322014000400013.json

4 S1139-76322013000200003.json

4 S0212-97282015000200009.json

4 S0212-97282014000200037.json

4 S1139-76322009000400002.json

4 S0211-69952016000600510.json

4 S1139-76322009000200005.json

5 S1699-695X2011000100005.json

5 S0211-69952017000100009.json

5 S1139-76322014000400014.json

5 S1699-695X2009000300004.json

5 S1139-76322010000500002.json

5 S0211-69952015000100002.json

5 S1135-76062002000200007.json

5 S1139-76322012000500004.json

5 S1139-76322016000400007.json

5 S0212-97282016000100003.json

Procesamiento completado en 219.91 segundos.

Archivos procesados: 1000

Total de tokens: 257595

Total sin stopwords: 148174

Total sin stopwords y stemmer: 148174

Promedio de tokens por archivo: 257.60

Promedio de tokens sin stopwords por archivo: 148.17

CONCLUSIÓN:
Creo que se ha quedado un trabajo funcional, es muy ineficiente, porque hace llamadas a memoria innecesarias, luego las funciones no son optimas ya que uso bucles que recorren continuamente todos los términos o toda la colección y eso habría que hacerlo 1 o 2 veces máximo, por lo que eso lo tengo que arreglar. Luego también debería de hacer limpieza de código, añadir más comentarios y explicaciones para que los nombres (que son algo confusos a veces) tengan algo más de convenio entre ellos. Para finalizar, la matriz dispersa tengo que simplificarla, para poder subirlo al github junto con todo el código.

Memoria de Mejoras Implementadas en el Sistema de Recuperación de Información

Introducción

El objetivo de estas mejoras es potenciar la efectividad y robustez del sistema de recuperación de información original. Se han incorporado siete mejoras que permiten optimizar tanto la recuperación de documentos como la relevancia de los resultados. Estas mejoras buscan abordar desafíos comunes en sistemas IR, como la escasez de términos en consultas breves, la ambigüedad semántica, y la necesidad de un ranking más robusto.

Descripción de las Mejoras y Justificación

1. Pseudo-realimentación por Relevancia (PRF)

Descripción:
Se ejecuta una primera recuperación utilizando la consulta original y se seleccionan los cinco documentos más relevantes. A partir de ellos, se extraen los cinco términos más frecuentes (totalizando 25 términos) que se añaden a la consulta para una segunda recuperación.
Justificación:
Esta técnica mejora la cobertura de la consulta al incluir términos contextualmente relevantes que podrían haber sido omitidos inicialmente, lo que incrementa la probabilidad de recuperar documentos pertinentes.

2. Expansión de Consulta

Descripción:
Se implementa la expansión de la consulta mediante un pequeño diccionario de sinónimos. Por ejemplo, términos como "ciencia" se amplían a "conocimiento" y "estudio".
Justificación:
La expansión de consulta ayuda a superar la limitación de vocabulario (vocabulary

mismatch) entre la consulta y los documentos, mejorando así la recuperación de información relevante.

3. Filtrado por Categorías

Descripción:
Una vez obtenidos los documentos relevantes, se extraen las categorías (por ejemplo, keywords presentes en el XML) y se permite al usuario filtrar los resultados según la categoría de interés.
Justificación:
El filtrado por categorías incrementa la precisión del sistema, permitiendo al usuario refinar los resultados y centrar la búsqueda en un ámbito temático específico.

4. Uso de BM25 para el Cálculo de Similitud

Descripción:
Se incorpora la métrica BM25, que ajusta la puntuación de relevancia basándose en la frecuencia del término, la longitud del documento y parámetros específicos (k1 y b).
Justificación:
BM25 es ampliamente reconocido por su rendimiento superior en comparación con la métrica TF-IDF tradicional. Su capacidad para manejar la variabilidad en la longitud de documentos y la distribución de términos lo hace ideal para sistemas IR modernos.

5. Indexación con Semántica Latente (LSI)

Descripción:
Se aplica la descomposición en valores singulares (SVD) a la matriz documento-término para reducir la dimensionalidad, extrayendo la "semántica" subyacente en los textos.
Justificación:
LSI permite capturar relaciones latentes entre términos y documentos, reduciendo el ruido y mejorando la recuperación en casos donde las coincidencias léxicas directas son insuficientes.

6. Uso de Otros Tokenizadores (Tokenización a Nivel de Sub-palabra)

Descripción:
Se incorpora la opción de utilizar tokenizadores preentrenados (por ejemplo, los basados en modelos BERT) que segmentan el texto a nivel de sub-palabra.
Justificación:
Esta estrategia mejora el manejo de palabras fuera de vocabulario y variaciones morfológicas, lo que resulta especialmente útil en lenguajes con alta inflexión como el español.

7. Codificación de Documentos con Modelos del Lenguaje (Embeddings)

Descripción:
Se integra una vía de búsqueda en la que tanto documentos como consultas se codifican utilizando modelos de sentence transformers. La similitud se mide mediante la distancia del coseno entre los vectores generados.
Justificación:
El uso de embeddings permite capturar la similitud semántica de forma más robusta que

los métodos tradicionales basados en frecuencias, facilitando la identificación de documentos relevantes incluso cuando la coincidencia lexical es limitada.

Conclusión

La integración de estas mejoras responde a la necesidad de contar con un sistema de recuperación de información más inteligente y adaptable. Cada técnica ha sido seleccionada por su capacidad para abordar problemas específicos:

- **PRF** y **expansión de consulta** mejoran la cobertura y relevancia al enriquecer la consulta original.
- **Filtrado por categorías** y **BM25** optimizan la precisión y el ranking de los resultados.
- **LSI** y **modelos de embeddings** permiten capturar relaciones semánticas complejas que los enfoques tradicionales no abordan.
- Finalmente, la opción de **tokenización a nivel de sub-palabra** mejora el procesamiento del lenguaje, manejando eficazmente la variabilidad lingüística.

Estas mejoras en conjunto elevan el rendimiento del sistema y lo acercan a estándares modernos en recuperación de información, ofreciendo resultados más precisos y relevantes para el usuario.