Informe del sistema de recuperación tradicional MiniTREC

Jaime Ruiz-Borau Vizárraga
546751

Alberto Sabater Bailón
546297

1. Arquitectura del software desarrollado

El siguiente diagrama de clases representa el diseño arquitectural del software desarrollado para el sistema de recuperación tradicional:

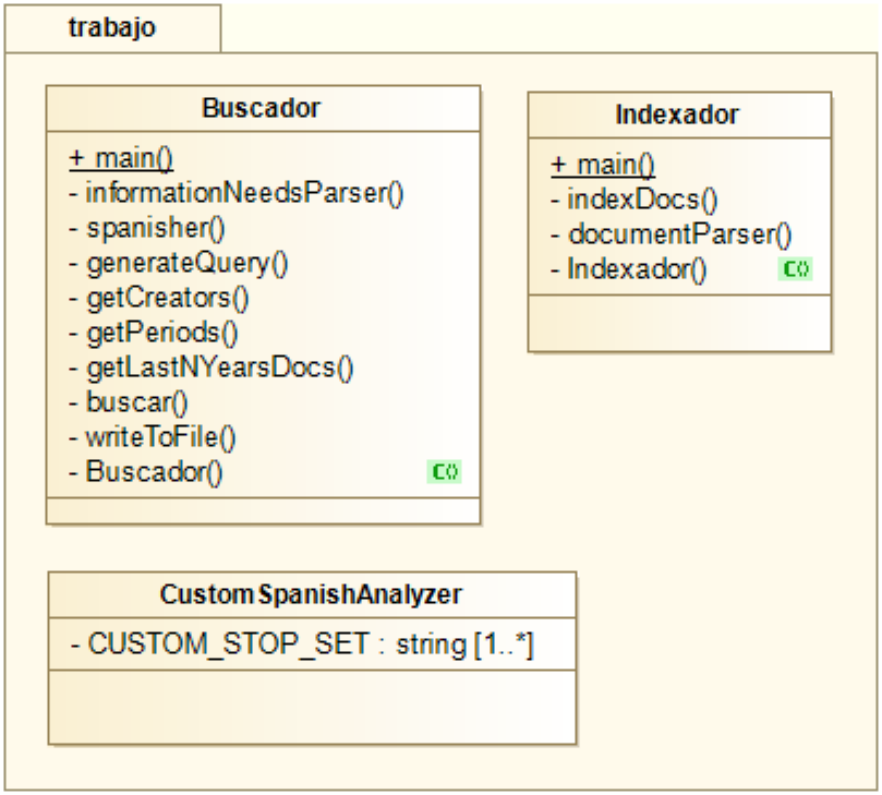

Figura 1: Diagrama de clases y paquetes del sistema

Como se puede apreciar en el diagrama, todas las clases se alojan en el paquete **trabajo**. Los nombres de las principales clases se explican por sí mismos: la clase *Indexador* es la que posee métodos para indexar una colección de documentos especificada por parámetro tal y como solicita el enunciado; y la clase *Buscador* es la que se encarga de realizar las consultas de las necesidades de información en base al índice creado por la clase Indexador.

Adicionalmente existe una tercera clase, *CustomSpanishAnalyzer*, que extiende a la clase de Lucene **Analyzer**. Reimplementa el método *createComponents*, que permite modificar las StopWords que suprimirá el analizador a la hora de lematizar una consulta.

2. Técnicas empleadas

2.1. Técnicas empleadas en el Indexador

Dada la naturaleza de la colección de documentos del Zaguán, se ha optado por realizar un análisis de las etiquetas de los documentos extrayendo la información de todas ellas siguiendo el estándar de la iniciativa Open Archives **(Open Archives Initiate)**. Dicho estándar establece una serie de etiquetas en documentos XML: **Title** *(título)*, **Creator** *(autor)*, **Subject** *(materia)*, **Description** *(descripción)*, **Publisher** *(editor)*, **Contributor** *(asistente)*, **Date** *(fecha)*, **Type** *(tipo)*, **Format** *(formato)*, **Identifier** *(identificador)*, **Source** *(fuente)*, **Language** *(idioma)*, **Relation** *(relacion)*, **Coverage** *(cobertura)* y **Rights** *(derechos)*.

Aunque varios de ellos no aparecen en la colección de documentos del Zaguán, se implementaron igualmente su análisis y extracción en el caso de que en el futuro se llegasen a incluir en los documentos del Zaguán (mayor modularidad futura).

Para cada etiqueta descrita anteriormente, simplemente se crearon los campos asociados con el mismo nombre y lo siguientes tipos:

- **TextField** - Title, Creator, Subject, Description, Publisher y Contributor
- **IntField** - Date
- **StringField** - Type, Format,

3. Puntos tratados

4. Otros puntos