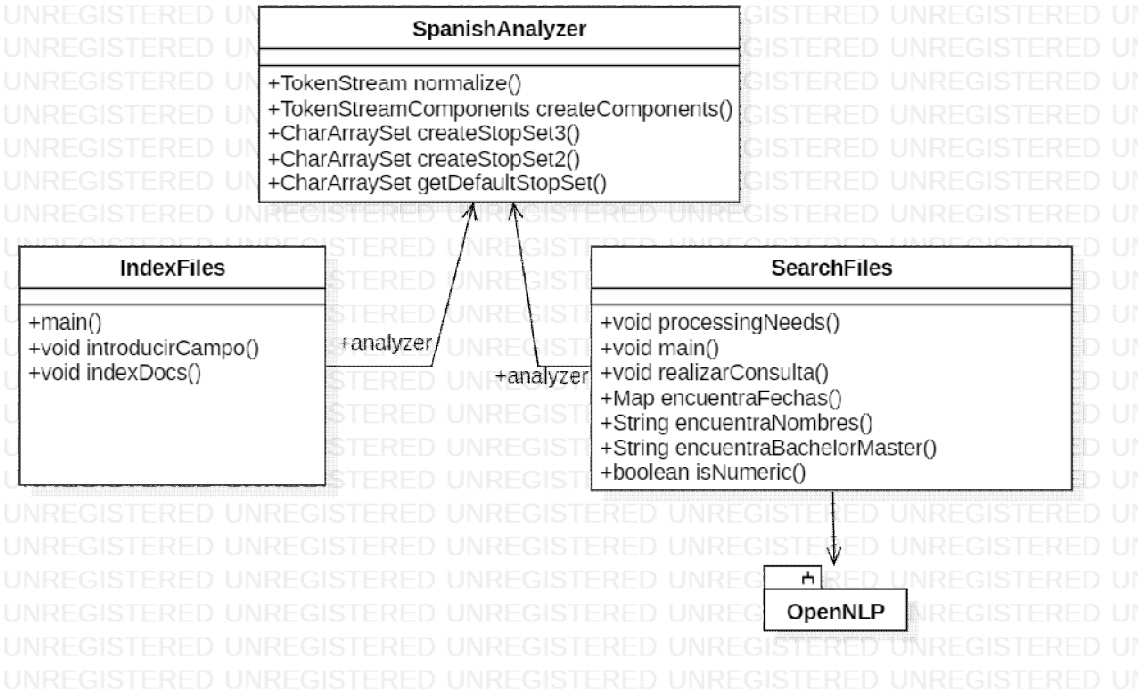
Trabajo 1 –Sistema de recuperación tradicional

Recuperación de Información

Equipo 16: Francisco Ferraz – Guillermo Cruz

Arquitectura de Software

Sobre el diseño de la arquitectura de software, se ha decidido adherirse al modelo presentado en las prácticas y modificarlo lo menos posible, para poder mantener la sencillez de la misma a pesar de alargar el código de cada clase.



Indexación

Para la indexación de los documentos se ha realizado una modificación del código utilizado en las prácticas de la asignatura, que utiliza la librería Lucene para operar. En concreto, se han limpiado las cadenas de texto originales, dejándolas en minúsculas y convirtiendo las fechas de formato AAAA-MM-DD (Año, Mes y Día, respectivamente) o AAAA a AAAAMMDD. Adicionalmente, se ha utilizado la clase SpanishAnalyzer (modificada de las prácticas por nosotros).

Índices creados:

Title: Título del documento.

Identifier: Identificador en la base de datos.

Subject: Temas que se tratan o palabras principales.

Creator: Creador, alumno que desarrolla el documento.

Publisher: Publicador del documento (en este caso casi siempre Unizar).

Date: Fecha (individual o rango)

Type: Tipo (TFG o TFM)

Description: Descripción del tema estudiado.

Language: Idioma en el que se desarrolla el documento.

Parseo de consultas

Para el parseo de consultas se ha utilizado una mezcla de uso de redes neuronales con una serie de medidas más ad-hoc desarrolladas por nosotros. Por partes:

Para el reconocimiento de nombres propios, la red neuronal proporcionada desde Apache OpenNLP obtenía muy malos resultados. La mayoría de nombres españoles que se han probado no eran reconocidos y además tenía problemas con el formato de texto a la hora de distinguir ciertas separaciones de palabras y conjuntos nombre-apellido. Por tanto, e han juntado una serie de documentos de nombres sacados de internet para hacer un fichero TXT, con el cual hacen búsquedas de coincidencias de la cadena original con cualquiera de las palabras del texto. Esto sirve de suplemento para OpenNLP, en los casos en los que no reconoce un nombre.

Para el reconocimiento de fechas, la red neuronal también daba problemas similares, por lo que hemos creado un método a parte que utiliza la clase HashMap para reconocer intervalos o fechas sueltas. Además, hacemos un pequeño post-procesado que nos permite convertir una frase como "...posterior al 2015" en un rango de 2015 a 2019. Solo funciona con algunas palabras concretas, pero ayuda en ciertos casos.

Para el procesado de temas, hemos utilizado Apache OpenNLP. Concretamente, el modelo *es-pos-maxent*, que es un Part-Of-Speech Tagger (Etiquetador de Partes del Lenguaje). Este modelo clasifica las palabras en una frase de manera morfológica y sintáctica, de tal modo que ayuda a reconocer los elementos en una frase que pueden corresponderse con parte del título, la descripción o los temas del documento.

Ranking

El ranking utilizado finalmente es el probabilista que utiliza Lucene por defecto. Se han hecho algunas pruebas con el modelo vectorial, pero los resultados no eran especialmente más satisfactorios.

Además, gracias a la red neuronal, se han podido identificar nombres comunes acompañados de adjetivos, y dichos conjuntos de un nombre común con un adjetivo, se han buscado conjuntamente con consultas del tipo 'PhraseQuery', por ejemplo: description:"enfermedades oculares". Este tipo de consultas, se repetían varias veces, para que los documentos que coincidieran subieran en el ranking y así estar mejor valorados.

Resultados

De forma general, nuestro algoritmo resulta eficiente a la hora de encontrar documentos que sigan las necesidades de información requeridas. En concreto, hace un muy buen trabajo encontrando nombres propios, y no se han detectado fallos respecto a consultas de fechas. Si bien esto es cierto, peca de dar excesiva importancia a los nombres en la puntuación final y devuelve cantidades muy elevadas de documentos para cada necesidad, pero gracias a la técnica de realizar consultas de nombres comunes con adjetivos a la vez, los documentos devueltos en las primeras posiciones sí que coinciden bastante con la necesidad de información presentada.