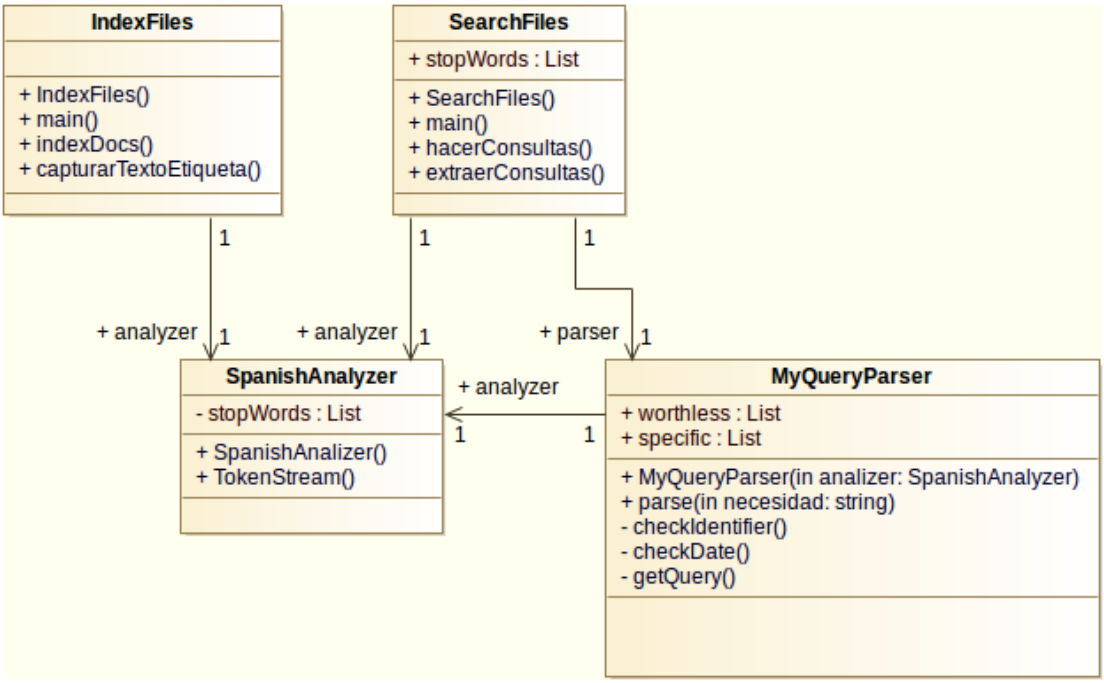
6-11-2016

Memoria MiniTREC: Sistema Recuperación tradicional

Recuperación de información

Jorge Sanz (680182) y Pablo Viñuales (679609)
UNIVERSIDAD DE ZARAGOZA

ARQUITECTURA DEL SISTEMA DESARROLLADO



El sistema está compuesto por 2 módulos, el módulo de búsquedas y el módulo de indexación. El módulo de indexación está compuesto por la clase IndexFiles que crea los índices correspondientes.

El módulo de búsqueda está compuesto por las clases MyQueryParser, que recibe un String con la consulta y lo transforma a una Query con la cual se realiza la búsqueda; y la clase SearchFiles que extrae las consultas del fichero XML y se las pasa al parser. Estas consultas tras ser parseadas son utilizadas para realizar las búsquedas sobre los índices especificados. Escribe los resultados obtenidos en un fichero de salida.

PROCESO DE INDEXACIÓN

Se han indexado documentos en formato XML que corresponden a información referente a los trabajos de fin de carrera, trabajos de fin de grado y tesis de la Universidad de Zaragoza que se encuentran en el repositorio Zaguan.

Para la indexación, se ha utilizado el Spanish Analyzer contenido en la librería de Lucene, ya que tanto las necesidades de información como los ficheros extraídos del repositorio Zaguan están en idioma Castellano y se considera que deben ser procesados con este analizador para reducir el espacio de almacenamiento y el coste de cálculo.

Los ficheros XML del repositorio Zaguan sobre los cuales hay que realizar las búsquedas, tienen una serie de etiquetas de las que solo hemos considerado útiles las siguientes:

- <dc:title>: que contiene el título del trabajo al que se hace referencia.
- <dc:identifier>: determina el tipo de trabajo (TFG, TFM, PFC o tesis).
- <dc:description>: que contiene una descripción del trabajo al que se hace referencia.
- <dc:creator> : que contiene el nombre del creador o creadores del trabajo.
- <dc:date> : que contiene la fecha de publicación del trabajo.

PROCESO DE PARSEO DE CONSULTAS

El fichero XML de necesidades de información pasado como parámetro a la clase de búsqueda SearchFiles es parseado para poder extraer las necesidades contenidas en él. Una vez extraídas y separadas, son pasadas de una en una al método "parse" del objeto MyQueryParser que realiza el parseo.

Lo primero que se realiza es la tokenización, normalización, stemming y eliminación de stopwords de las necesidades de información, mediante el método tokenStream del Analyzer. A partir de esta lista se comprueban sus características mediante una serie de métodos para así crear una consulta booleana que agrupe dichas caracteristicas. Los métodos ejecutados son:

- checkIdentifier: que detecta el tipo de trabajo solicitado.
- checkDate: que a partir de una fecha o un intervalo detecta todos los trabajos que se han publicado en esa época, para diferenciar las fechas de publicación de otras se han considerado únicamente aquellas que se encuentran próximas al término "trabajo" o "publicación". También se han considerado solo aquellas fechas que podrían hacer referencia a la publicación de un trabajo (1980 – 2070).
- getQuery: devuelve una BooleanQuery compuesta por TermQuerys para cada uno de los términos de la lista anterior que no se encuentren en la lista Worthless (palabras que no aportan significado a la consulta).

El resultado del parseo es una BooleanQuery que agrupa las consultas anteriores. Las consultas acerca de los campos título, descripción y creador tienen un peso que es respectivamente 2, 1 y 3.

Una vez parseada la consulta se ejecuta el método principal de la clase SearchFile, que busca sobre los índices anteriormente creados y escribe en un fichero los resultados devueltos.

CÁLCULO DEL RANKING

Los resultados obtenidos para cada necesidad de información, son escritos ordenadamente en un fichero según el score obtenido. Así, los ficheros con mayor score, es decir, con mayores apariciones y más importantes de los términos que aparecen en la necesidad de información según los pesos especificados anteriormente, aparecen antes en el fichero.

RESULTADOS OBTENIDOS

Para la entrada del fichero XML de necesidades de información, se ha generado un fichero de salida en el cual se han escrito los ficheros resultantes de realizar la búsqueda con el formato que se especifica en el enunciado de este trabajo.

- Para la primera consulta, se han encontrado los ficheros correspondientes y han sido ordenados correctamente.
- Para la segunda consulta, también se han encontrado documentos relacionados con el tema solicitado, pero no en el rango de fechas establecido.
- Para la tercera consulta, se han encontrados ficheros de trabajos de algoritmos relacionados con la geometría.
- Para la cuarta consulta, se han encontrado trabajos académicos relacionados con la robótica publicados a partir del año 2005 pero tan solo uno de los encontrados (el primero) pertenece a algún miembro de la familia Martínez.
- Para la quinta consulta, encuentra trabajos relacionados con la inteligencia artificial en los videojuegos en los últimos 5 años.

Ya que, se ha tenido que especificar en el fichero de salida todos los fichero resultado, los más bajos del ranking para cada necesidad probablemente no estén tan relacionados con la consulta.