Trabajo de la asignatura de Recuperación de Información: MiniTREC

08 de noviembre de 2018

Víctor Labrador Ortiga, 701658

David Márquez Calavia, 700940

Arquitectura del software desarrollado

La clase IndexFiles genera los índices invertidos sobre los documentos y los almacena en un directorio a especificar.

SearchFiles es la clase encargada de leer las consultas de un fichero a especificar por el usuario en formato xml, realiza transformaciones sobre las mismas, las cuales se explicarán más adelante, y escribe en un fichero, también introducido por el usuario, y el resultado de las búsquedas, empleando los índices generados anteriormente.

Las clases de analizadores se emplean tanto en IndexFiles como en SearchFiles para analizar los textos en cuestión.

Técnicas utilizadas en el proceso de indexación

Para realizar la indexación se ha creado un Analizador para texto, en el cual se aplican las siguientes transformaciones:

- StandardFilter: eliminar signos de puntuación
- LowerCaseFilter: transformar el texto a minúsculas
- StopFilter: para eliminar las palabras en el fichero de stopwords, el cual se ha creado a partir de varios ficheros preexistentes.
- SpanishLightStemFilter: para lematizar las palabras

También se ha empleado el modelo probabilístico Okapi BM25, una función de ranking para asignar relevancia a ciertos documentos y ordenarlos en función a ello.

Se ha creado otro analizador a parte, para el caso de los nombres, donde se aplican estas transformaciones:

- StandardFilter: eliminar signos de puntuación
- LowerCaseFilter: transformar el texto a minúsculas

No se han realizado tantas transformaciones ya que no tiene sentido lematizar un nombre propio o eliminar stopwords.

Índices creados

Etiqueta XML	Nombre de campo	Field asociado
dc:title	title	TextField
dc:subject	subject	TextField
dc:type	type	StringField
dc:description	description	TextField
dc:creator	creator	TextField
dc:date	date	TextField

Técnicas utilizadas en el parseo de las consultas

- Date:
  (publicados entre |periodo | a partir de)(?<anyo>\\d\\d\\d\\d)
  (y |-)(?<anyo>\\d\\d\\d\\d)
  (los últimos)(?<anyo>\\d)
  De esta manera se saca el rango de años sobre el que le interesa al usuario realizar su consulta, en caso de no especificarse se realiza en todos los años.
- Type:
  "master| tesis de fin de master| tesis de master|trabajo fin de master |trabajo de doctorado|doctora", para el caso de masterthesis
  "tfg |trabajo fin de grado", para el caso de bachelorthesis.

Este parseo a priori no era necesario ya que en las consultas se piden trabajos de master o de fin de grado indistintamente, pero se ha añadido de todos modos.

- Creator:

  Se detecta en el campo creator toda palabra que empiece por mayúscula seguido de letras en minúscula, permitiendo también nombres compuestos mediante la siguiente expresión regular: "(?<nombre>[Á-ÚA-Z][a-zá-ú]+)".
  Después, se compara el resultado obtenido con un fichero de nombres creado previamente que contiene los nombres más comunes en la lengua castellana. Si se encuentra una coincidencia, se guarda para realizar la consulta más adelante.

- Resto de campos:

  Para los campos de title, subject y description se ha cogido el texto entero, pasándolo por la clase Analyzer para realizar las transformaciones especificadas anteriormente.

Algoritmo elegido para el cálculo de ranking

Se ha empleado BoostQuery para indicar el orden de preferencia en los resultados obtenidos, siendo creator el más importante, seguido de subject, title, date y description.

Los pesos asignados son 5 para creator, 1.4 para subject, 1.2 para title, 1 para date y 0.3 para description.

Breve comentario sobre los resultados obtenidos

La comprobación de la precisión de los resultados se ha realizado leyendo los 10 primeros documentos recuperados según el algoritmo de ranking, y corroborando que tienen importancia para cada búsqueda reflejada en el fichero xml.

No se ha podido conseguir la recuperación de los documentes más importantes en todos los casos debido a la dificultad de interpretar el lenguaje, por lo que aquellas consultas que preferentemente el nombre de pila de la autora fuera María o aquella que pedía estudios para profesorado de ESO y bachillerato pero NO FP no son tan precisas como podrían ser.