Prácticas Recuperación de Información.
Grado en Ingeniería Informática.

P2. Ejercicio 1

Buscador para un sistema de desktop search

Estudie y pruebe el código
http://lucene.apache.org/core/4_10_3/demo/src-html/org/apache/lucene/demo/SearchFiles.html

P2. Ejercicio 2.

Se trata de construir un lector/buscador para índices (IndexInspector) con las siguientes opciones. En los resultados de búsquedas (query, multiquery y progquery) debe aparecer siempre el doc id, el título y el score de los documentos que satisfacen la consulta. Probar IndexInspector con los índices creados en la práctica 1.

-index indexfile (indexfile es la ruta donde está el índice que se procesa)
-out indexfile (crea un índice con los documentos que satisfacen la -query, -multiquery o -progquery; o con los documentos especificados en la opción -docs; o con los documentos mostrados en las opciones -docswithtermFreq -docswithtermFreq2, y lo almacena en indexfile)
-query fld1 “query” (muestra los resultados de la query “query”, que es cualquier query aceptada por el query parser, lanzada sobre el campo fld1). Normalmente cualquier intérprete de comandos interpreta un argumento entre comillas como un argúmento único, por tanto si se quiere que el argumento contenga comillas, están deben ir *escaped,* compruebe que éste es el comportamiento en su intérprete (el de Eclipse u otro). Por tanto para hacer una Phrase Query debiera indicar -query field “\”New York\””.
-multiquery fld1 “query1” fld2 “query2” … (muestra los resultados de las queries “queryi”, que son cualquier query aceptada por el query parser, lanzadas sobre los campos fldi. Usar MultiFieldQueryParser). Mirar la documentación del método parse del MultiFieldQueryParser si quiere cambiar el comportamiento OR por defecto entre cláusula (BooleanClause.Ocurr)
-progquery fld1 -and term1 term2 … -or termi termj … -not termp termq … - (muestra los resultados de la query construida programáticamente y lanzada sobre el campo fld1).
-write file (vuelca los contenidos del índice en formato texto plano sobre el fichero file)
-docs i j (muestra los contenidos, es decir todos los campos, de los documentos del rango i a j)
-termswithdocFreq field n1 n2 (muestra los términos del campo fld con docFreq, df, mayor igual que n1 y menor igual que n2)
-docswithtermFreq field n1 n2 (para cada término del campo field muestra los documentos en los que el término -term, field- tiene frecuencia de documento, tf, mayor o igual que n1 y menor que n2)
-docswithtermFreq2 term field n1 no (muestra los documentos en los que el término -term, field- tiene frecuencia de documento, tf, mayor o igual que n1 y menor igual que n2)
-topterms field n (para el campo field devuelve el top n de términos ordenados por tf x idf)
tf x idf (t, d) =(1 + log (tf, d)) x log (N / df(t))
-bottomterms field n (para el campo field devuelve los últimos n de términos ordenados por tf x idf)
-midterms field n1 n2 (para el campo field devuelve los términos situados en las posiciones mayor igual que n1 y menor igual que n2, en la ordenación por tf x idf).
-topterms2 field doc n (para el campo field y el documento doc devuelve el top n de términos ordenados por tf x idf)
tf x idf (t, d) =(1 + log (tf, d)) x log (N / df(t))
-bottomterms2 field doc n (para el campo field y documento doc devuelve los últimos n de términos ordenados por tf x idf)
-midterms2 field doc n1 n2 (para el campo field y documento doc devuelve los términos situados en

las posiciones mayor igual que n1 y menor igual que n2, en la ordenación por tf x idf).

Entrega P2

-Se sube al repositorio SVN antes de la fecha límite indicada
-Se crea una carpeta P2 y se sube un archivo con el nombre IndexInspector.java, y con el nombre de la clase principal coincidente con el del archivo. SOLAMENTE puede subirse ese archivo por lo que las clase parser y cualquier otra necesaria deben incluirse en ese archivo con ese nombre. SOLO UNO DE LOS ALUMNOS de la pareja de prácticas puede subir la práctica, si la suben los dos, los scripts de bajada y procesamiento detectarán copia y se invalidará la práctica. Si la subida no se hace en las condiciones (nombre de carpeta y archivo, archivo único) establecidas, los scripts de procesamiento tampoco detectarán la práctica como correcta.
-LAS PRÁCTICAS SON POR PAREJAS.
-EN EL CASO DE COPIA DE PRÁCTICAS, LOS ALUMNOS IMPLICADOS PERDERÁN LA NOTA TOTAL DE PRÁCTICAS.
-Además de la entrega habrá una revisión in situ de las prácticas donde se pedirán cambios que deberán implementarse en el aula de prácticas tanto para el ejercicio 1 como para el ejercicio 2. Cualquier miembro de la pareja de prácticas debe responder a cualquier aspecto de la defensa, también se puede pedir que cada uno implemente una variante distinta. Por tanto aunque el trabajo es en equipo cada miembro debe conocer al detalle lo realizado por el compañero.