Prácticas Recuperación de Información.
Grado en Ingeniería Informática.

P1. Ejercicio 1.

Descargar e indexar la colección Reuters 21578
http://kdd.ics.uci.edu//databases/reuters21578/reuters21578.html

Tener en cuenta:

-Los archivos a indexar son los *.sgm
-El tag REUTERS identifica el principio y final de cada artículo.
-Debéis indexar los campos TOPICS, TITLE, DATELINE, BODY y DATE.
-Cada tópico individual en TOPICS está identificado por un tag D.
-Si el BODY finaliza con "Reuter&#3;" ó "REUTER&#3;" (boiler plate) el parser debe eliminar estos textos para no indexarlos.

Se os proporciona un parser de la colección Reuters que ya se ocupa de la extracción de los campos y del boiler plate.

Indexación del campo DATE:

En primer lugar se necesita un cambio mínimo en el parser Reuters21578Parser.java para procesar el campo DATE de la colección Reuters
La clase SimpleDateFormat de Java permite parsear y formatear fechas. En nuestro caso basta usar el constructor de la clase pasándole un string con el formato de la fecha según el formato del campo DATE de la colección Reuters
Invocando el método parse del objeto creado en el paso anterior sobre el string obtenido del campo DATE se obtiene un objeto de la clase Date de Java
Con el método dateToString de la clase DateTools de Lucene se convierte el objeto Date de Java en un string para almacenar en el campo correspondiente del índice

Argumentos que debe aceptar el indexador y acciones en su caso:

-openmode openmode (el open mode será append, create, o create_or_append)
-index pathname (carpeta que contiene o contendrá el índice)
-files pathname (carpeta que contiene los archivos *.sgm)
-onlyfiles entero (entero en el rango 0 a 21, que indicará que sólo se indexen los artículos Reuters del correspondiente archivo *.sgm)
-addsgmfile (indica que se indexe con un campo adicional que contiene el nombre del archivo *.sgm donde se encuentra el artículo Reuters)
-delete txt field (borrará los documentos que contienen el término txt en el campo field). Vea la documentación de la clase Term para construir un término

P1. Ejercicio 2

Indexador para un sistema de desktop search

Estudie y pruebe el código
http://lucene.apache.org/core/4_0_0/demo/src-html/org/apache/lucene/demo/IndexFiles.html

Entrega P1

-Se sube al repositorio SVN antes de la fecha límite indicada
-Se crea una carpeta P1 y se sube un archivo con el nombre Reuters21578Indexer.java, y con el nombre de la clase principal coincidente con el del archivo. SOLAMENTE puede subirse ese archivo por lo que la clase Reuters21578Parser con la modificación necesaria debe incluirse en ese archivo.
-LOS NOMBRES DE CARPETA Y ARCHIVO TIENEN QUE SER EXACTAMENTE ESOS.
-En el ejercicio 2 no hay que subir nada
-LAS PRÁCTICAS SON INDIVIDUALES. EN EL CASO DE COPIA DE PRÁCTICAS, LOS ALUMNOS IMPLICADOS PERDERÁN LA NOTA DE PRÁCTICAS.
-Además de la entrega habrá una revisión in situ de las prácticas donde se pedirán cambios que deberán implementarse en clase de prácticas tanto para el ejercicio 1 como para el ejercicio 2.