INSTITUTO TECNOLÓGICO DE COSTA RICA II SEMESTRE 2016
INGENIERÍA EN COMPUTACIÓN
IC-8020 ELECTIVA TEXTO

TAREA PROGRAMADA 2

1. Introducción

La tarea consiste en indexar y buscar una colección de documentos usando Lucene de Apache. La colección usada será la colección Reuters21578 que se les adjunta. Esta es una colección que ha sido muy usada en investigación en information retrieval y machine learning. Consiste de 21578 artículos aparecidos en el servicio de noticias Reuters en 1987. Los artículos fueron indexados y categorizados por el personal de Reuters Ltd. en 1987.

La colección consiste de 22 archivos XML, cada uno de ellos, menos el último, contienen 1000 articulos. Los artículos son representados por medio del elemento <REUTERS>. Dentro de ese elemento aparecen los siguientes elementos con la información indicada:

<DATE>	Fecha y hora de la noticia. Formato dd-MMM-YYYY hh.mm.ss.cc
<TOPICS>	Categorías asignadas a un articulo. Por ejemplo: grain, sugar, etc. Cada categoría se encuentra dentro de un elemento <D>. Puede haber más de una.
<PLACES>	Lugares geográficos a los que se refiere la noticia. Por ejemplo: usa, uk, etc. Cada lugar se encuentra dentro de un elemeno <D>. Puede haber más de uno.
<PEOPLE>	Personas mencionadas en la noticia. Por ejemplo: reagan, volcker, etc. Cada persona se encuentra dentro de un elemeno <D>. Puede haber más de una.
<ORGS>	Organizaciones mencionadas en la noticia. Por ejemplo: oecd, imf, etc. Cada organización se encuentra dentro de un elemeno <D>. Puede haber más de una.
<EXCHANGES>	Bolsa de valores mencionadas en la noticia. Por ejemplo: nysec, nasdaq, etc. Cada bolsa se encuentra dentro de un elemeno <D>. Puede haber más de una.
<COMPANIES>	Aunque interesante este campo no tiene datos. Ignorar.
<UNKNOWN> <MKNOTE>	Campos sin interés. Deben ser ignorados.
<TEXT>	Contiene la noticia. Desglosada de la siguiente manera:
<TITLE>	Título de la noticia
<AUTHOR>	Autor o autores de la noticia. No están separados por elementos <D>.
<DATELINE>	Lugar y fecha dentro de la noticia. Por ejemplo: Quito, march 18

<BODY>	Cuerpo de la noticia

2. Tabla de estado del proyecto

A continuación se muestra la tabla donde se evidencia el estado final del proyecto

Etapa	% de complet.	Comentario o aclaración
Primer indexado: **indexar reut2-000.xml -- reut2-010.xml**	100%	
Resultados para primer indexado		
Prueba 1 File: reut2-004.xml <OLDID="9516" NEWID="4603"> File: reut2-005.xml <OLDID="10384" NEWID="5471"> File: reut2-007.xml <OLDID="12037" NEWID="7124"> File: reut2-001.xml <OLDID="16532" NEWID="1212"> File: reut2-001.xml <OLDID="18133" NEWID="1715"> File: reut2-001.xml <OLDID="18298" NEWID="1880"> File: reut2-002.xml <OLDID="18885" NEWID="2467"> File: reut2-002.xml <OLDID="19372" NEWID="2954"> File: reut2-005.xml <OLDID="9915" NEWID="5002"> File: reut2-005.xml <OLDID="10047" NEWID="5134"> File: reut2-008.xml <OLDID="14424" NEWID="8903">		Los resultados en azul corresponde a los documentos esperados en la prueba. El resto son noticias nuevas recuperadas.
Prueba 2 File: reut2-005.xml <OLDID="10202" NEWID="5289"> File: reut2-006.xml <OLDID="11027" NEWID="6114"> File: reut2-001.xml <OLDID="16566" NEWID="1246"> File: reut2-006.xml <OLDID="11803" NEWID="6890"> File: reut2-007.xml <OLDID="13438" NEWID="7917"> File: reut2-008.xml <OLDID="14226" NEWID="8705"> File: reut2-003.xml <OLDID="8894" NEWID="3981"> File: reut2-009.xml <OLDID="15041" NEWID="9521">		Idem
Prueba 3 File: reut2-001.xml <OLDID="16566" NEWID="1246">		Idem
Prueba 4 File: reut2-010.xml <OLDID="15579" NEWID="10058"> File: reut2-000.xml <OLDID="12652" NEWID="469"> File: reut2-007.xml <OLDID="12419" NEWID="7506"> File: reut2-004.xml <OLDID="9746" NEWID="4833"> File: reut2-003.xml <OLDID="8402" NEWID="3489"> File: reut2-003.xml <OLDID="8468" NEWID="3555"> File: reut2-006.xml <OLDID="11314" NEWID="6401"> File: reut2-009.xml <OLDID="14662" NEWID="9141"> File: reut2-009.xml <OLDID="14733" NEWID="9212"> File: reut2-001.xml <OLDID="17997" NEWID="1579"> File: reut2-004.xml <OLDID="9006" NEWID="4093">		Idem

Segundo indexado: agregar reut2-011.xml -- reut2-015.xml	100%	
¿No reindexa todo sino que es incremental? SI/NO	NO	
Resultados para segundo indexado		
Prueba 1 File: reut2-004.xml <OLDID="9516" NEWID="4603"> File: reut2-005.xml <OLDID="10384" NEWID="5471"> File: reut2-007.xml <OLDID="12037" NEWID="7124"> File: reut2-015.xml <OLDID="4539" NEWID="15357"> File: reut2-001.xml <OLDID="16532" NEWID="1212"> File: reut2-001.xml <OLDID="18133" NEWID="1715"> File: reut2-001.xml <OLDID="18298" NEWID="1880"> File: reut2-002.xml <OLDID="18885" NEWID="2467"> File: reut2-002.xml <OLDID="19372" NEWID="2954"> File: reut2-005.xml <OLDID="9915" NEWID="5002"> File: reut2-005.xml <OLDID="10047" NEWID="5134"> File: reut2-008.xml <OLDID="14424" NEWID="8903"> File: reut2-012.xml <OLDID="21546" NEWID="12689"> File: reut2-015.xml <OLDID="4181" NEWID="15198"> File: reut2-015.xml <OLDID="4902" NEWID="15725"> File: reut2-015.xml <OLDID="4914" NEWID="15737">		
Prueba 2 File: reut2-005.xml <OLDID="10202" NEWID="5289"> File: reut2-006.xml <OLDID="11027" NEWID="6114"> File: reut2-011.xml <OLDID="17597" NEWID="11536"> File: reut2-001.xml <OLDID="16566" NEWID="1246"> File: reut2-013.xml <OLDID="2296" NEWID="13313"> File: reut2-006.xml <OLDID="11803" NEWID="6890"> File: reut2-012.xml <OLDID="388" NEWID="12160"> File: reut2-007.xml <OLDID="13438" NEWID="7917"> File: reut2-008.xml <OLDID="14226" NEWID="8705"> File: reut2-003.xml <OLDID="8894" NEWID="3981"> File: reut2-009.xml <OLDID="15041" NEWID="9521"> File: reut2-015.xml <OLDID="4431" NEWID="15043">		
Prueba 3 File: reut2-001.xml <OLDID="16566" NEWID="1246">		
Prueba 4 File: reut2-010.xml <OLDID="15579" NEWID="10058"> File: reut2-000.xml <OLDID="12652" NEWID="469"> File: reut2-007.xml <OLDID="12419" NEWID="7506"> File: reut2-004.xml <OLDID="9746" NEWID="4833"> File: reut2-003.xml <OLDID="8402" NEWID="3489"> File: reut2-003.xml <OLDID="8468" NEWID="3555"> File: reut2-006.xml <OLDID="11314" NEWID="6401"> File: reut2-009.xml <OLDID="14662" NEWID="9141"> File: reut2-009.xml <OLDID="14733" NEWID="9212"> File: reut2-001.xml <OLDID="17997" NEWID="1579"> File: reut2-004.xml <OLDID="9006" NEWID="4093"> File: reut2-011.xml <OLDID="9" NEWID="11781"> File: reut2-012.xml <OLDID="1845" NEWID="12876">		

Tercer indexado: agregar reut2-016.xml -- reut2-021.xml	100%	
¿No reindexa todo sino que es incremental? SI/NO	NO	
Resultados para tercer indexado		
Prueba 1 File: reut2-004.xml <OLDID="9516" NEWID="4603"> File: reut2-005.xml <OLDID="10384" NEWID="5471"> File: reut2-007.xml <OLDID="12037" NEWID="7124"> File: reut2-015.xml <OLDID="4539" NEWID="15357"> File: reut2-016.xml <OLDID="814" NEWID="16279"> File: reut2-001.xml <OLDID="16532" NEWID="1212"> File: reut2-001.xml <OLDID="18133" NEWID="1715"> File: reut2-001.xml <OLDID="18298" NEWID="1880"> File: reut2-002.xml <OLDID="18885" NEWID="2467"> File: reut2-002.xml <OLDID="19372" NEWID="2954"> File: reut2-005.xml <OLDID="9915" NEWID="5002"> File: reut2-005.xml <OLDID="10047" NEWID="5134"> File: reut2-008.xml <OLDID="14424" NEWID="8903"> File: reut2-012.xml <OLDID="21546" NEWID="12689"> File: reut2-015.xml <OLDID="4181" NEWID="15198"> File: reut2-015.xml <OLDID="4902" NEWID="15725"> File: reut2-015.xml <OLDID="4914" NEWID="15737"> File: reut2-019.xml <OLDID="7223" NEWID="19387"> File: reut2-021.xml <OLDID="20259" NEWID="21177">		
Prueba 2 15 NOTICIAS RECUPERADAS File: reut2-005.xml <OLDID="10202" NEWID="5289"> File: reut2-019.xml <OLDID="7911" NEWID="19583"> File: reut2-006.xml <OLDID="11027" NEWID="6114"> File: reut2-011.xml <OLDID="17597" NEWID="11536"> File: reut2-001.xml <OLDID="16566" NEWID="1246"> File: reut2-013.xml <OLDID="2296" NEWID="13313"> File: reut2-006.xml <OLDID="11803" NEWID="6890"> File: reut2-012.xml <OLDID="388" NEWID="12160"> File: reut2-007.xml <OLDID="13438" NEWID="7917"> File: reut2-008.xml <OLDID="14226" NEWID="8705"> File: reut2-019.xml <OLDID="7098" NEWID="19262"> File: reut2-021.xml <OLDID="20013" NEWID="21423"> File: reut2-003.xml <OLDID="8894" NEWID="3981"> File: reut2-009.xml <OLDID="15041" NEWID="9521"> File: reut2-015.xml <OLDID="4431" NEWID="15043">		
Prueba 3 1 NOTICIAS RECUPERADAS File: reut2-001.xml <OLDID="16566" NEWID="1246">		
Prueba 4 14 NOTICIAS RECUPERADAS File: reut2-010.xml <OLDID="15579" NEWID="10058"> File: reut2-000.xml <OLDID="12652" NEWID="469"> File: reut2-007.xml <OLDID="12419" NEWID="7506"> File: reut2-004.xml <OLDID="9746" NEWID="4833"> File: reut2-003.xml <OLDID="8402" NEWID="3489"> File: reut2-003.xml <OLDID="8468" NEWID="3555"> File: reut2-006.xml <OLDID="11314" NEWID="6401"> File: reut2-009.xml <OLDID="14662" NEWID="9141"> File: reut2-009.xml <OLDID="14733" NEWID="9212">		

File: reut2-001.xml <OLDID="17997" NEWID="1579"> File: reut2-004.xml <OLDID="9006" NEWID="4093"> File: reut2-016.xml <OLDID="761" NEWID="16226"> File: reut2-011.xml <OLDID="9" NEWID="11781"> File: reut2-012.xml <OLDID="1845" NEWID="12876">		

3. Comentarios finales

El proyecto fue finalizado en un porcentaje altamente exitoso; sin embargo, existe problemas en la búsqueda en los noticias. El sistema no obtiene en ocaciones todos las noticias con el término que se está buscando.

En la pruebas se obtuvieron resultados exitosos en su mayoría y un poco ocurrencia de errores que no pudieron ser solventados.

Además hay problemas en el indexamiento y busqueda en fechas. El programa no permite realizar busquedas sobre rangos de fechas puesto que estas son indexadas como texto.