INSTITUTO TECNOLÓGICO DE COSTA RICA **I SEMESTRE 2017**
INGENIERÍA EN COMPUTACIÓN
IC-4302 BASES DE DATOS 2

TAREA PROGRAMADA 3

La tarea consiste en cargar una colección de documentos en una base de datos mongoDB. Dicha colección es llamada ***Reuters21578*** y 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. Cada artículo es representado por medio del elemento <REUTERS>. Dentro de ese elemento aparecen los siguientes elementos XML 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

Carga

La aplicación recibe como dato un directorio el cual contiene los archivos XML con los datos a cargar. Los datos se deben cargar en una colección con un nombre escogido por el usuario.

La colección mongoDB que almacene esos datos debe contener un documento por cada artículo de la colección ***Reuters21578***. Los elementos TOPICS, PLACES, PEOPLE, ORGS y EXCHANGES deben consistir de arreglos con los valores de los elementos D anidados. Se debe además extraer el valor del atributo NEWID del elemento REUTERS.

Indexado

Se deben crear los siguientes índices para facilitar el acceso a la información cargada.

- Crear un índice ascendente para cada uno de los campos de arreglos (TOPICS, PLACES, PEOPLE, ORGS y EXCHANGES).
- Crear un índice de texto para los campos TITLE y BODY.

Búsqueda

Usando la herramienta de comandos de mongoDB. Realizar las siguientes consultas. Para cada consulta se debe imprimir el identificador (NEWID) y el título (TITLE).

- sugar (en TOPICS) and indonesia (en PLACES)
- colombia and coffee (en BODY)

Operación adicional

Usar el mecanismo de mapReduce para contar la frecuencia de cada valor del campo PLACES. Esto es, si se tiene

```
{ "_id":1, ..., "places": [ "costa-rica", "el-salvador" ] }
{ "_id":2, ..., "places": [ "argentina", "el-salvador","usa" ] }
{ "_id":3, ..., "places": [ "usa", "colombia" ] }
```

se obtenga:

```
{ "_id" : "costa-rica", "value" : 1 }
{ "_id" : "el-salvador", "value" : 2 }
{ "_id" : "argentina", "value" : 1 }
{ "_id" : "usa", "value" : 2 }
{ "_id" : "colombia", "value" : 1 }
```

Consideraciones finales

Pueden usar cualquier lenguaje de programación. Se recomiendan Python y Java.

Investigar y usar herramientas para convertir XML en JSON. Un punto de partida es Xml to JSON parser-converter in Python. · GitHub.

Investigar mecanismos para cargar datos JSON en mongoDB.

La tarea puede ser realizada en grupos de dos personas.

La fecha límite de entrega es el lunes 11 de junio.