Bases de Datos Documentales - XML

Business & Finance, Cars & Transportation, Computers & Internet y Entertainment & Music, categorías de Yahoo Answer con las que se ha trabajado

Fecha: 31 de Marzo de 2022

Asignatura: Bases de Datos Avanzadas

Profesor: Julián Ruiz Fernández

Componentes: Jesús Baltasar Fernández
Carlos Gómez Fernández
Domenico Fasciano

Tabla de contenido

1. Resumen

El desarrollo del trabajo se basa en a partir de un archivo *XML*, el cual contiene toda la información relacionada con las preguntas realizadas en el portal *web Yahoo Answer*, entre los años 2005 y 2006, obtener la información de las preguntas que formen parte de las categorías seleccionadas, para su posterior carga en un motor de base de datos documental, en el cual realizar consultas sobre las misma para obtener información relevante.

2. Introducción

Para facilitar la comunicación entre los componentes del equipo en los avances en el desarrollo del trabajo se ha optado por la utilización de la plataforma *GitHub*, para controlar y mantener las versiones del trabajo en su desarrollo incremental.

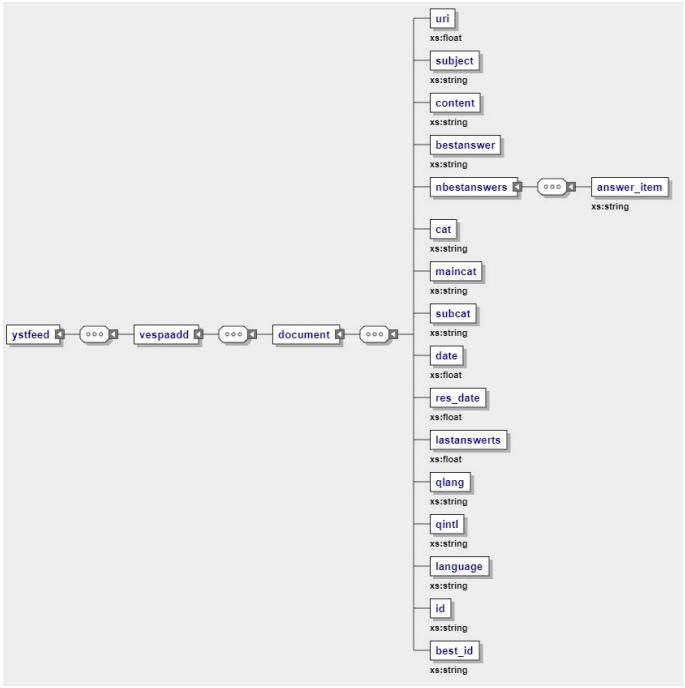
En cuanto al desarrollo del cuaderno de laboratorio se optó por la utilización de una de las plantillas proporcionadas en el enunciado y el cual se ha realizado haciendo uso de la herramienta *Overleaf*, la cual proporciona un editor del lenguaje *LaTeX*.

Uno de los primeros hitos importantes fue la selección de las herramientas y aplicaciones que se emplearán en el desarrollo del trabajo, entre las cuales cabe destacar *Eclipse*, para la extracción de la información de las categorías seleccionadas del *XML* mediante un programa en *Java*, *eXist-db*, para la carga de los archivos *XML* en la base de datos documental y *eXide*, para la realización de las consultas *XML* mediante el lenguaje de consultas *XQuery*.

3. Desarrollo

a. Procesamiento de datos

Como el archivo original tiene una extensión con la que es imposible trabajar, se optó por realizar un programa en *Java* haciendo uso del *IDE Eclipse,* el cual se ejecutó en tantas ocasiones como categorías se seleccionaron para el desarrollo del trabajo, obteniendo un total de cuatro archivos *XML*, con los cuales iniciaríamos el trabajo. A partir de uno de ellos, cualquiera, ya que la estructura que siguen es la misma, se generó el archivo *XSD*, el cual representa el esquema *XML* que sigue el archivo original. Posteriormente se generó también el archivo *XSLT*, a partir de un archivo *XSL*, el cual serviría de plantilla para un posible uso futuro. Finalmente se cargaron los archivos *XML* generados inicialmente en el motor de base de datos documental, teniendo que, uno de ellos dividirlo, debido a que su extensión excedía el límite admitido por el *software*.



b. Generación de archivos

Como se ha comentado anteriormente, los archivos *XML* en los que se almacenan las preguntas por categorías se obtuvieron mediante un programa *Java*. Por otra parte el archivo *XSD* se generó haciendo uso de una herramienta *on-line* proporcionada por el *Instituto Tecnológico de Massachusetts (MIT)* y finalmente el archivo *XSLT* se obtuvo haciendo uso de la herramienta *XMLmax*.

c. Bases de datos documental

Como hemos comentado en el *Lab-Book* del proyecto, hemos tenido bastantes dificultades para realizar la instalación de la base de datos documental, la cual es *eXist-db*, siéndonos imposible realizar su instalación en *Windows*, y optando por su instalación en una *Virtual Machine* con sistema operativo *Linux*.

Tras su instalación, la cual requiere de *Java*, se deberán seguir los siguientes pasos para poner en funcionamiento el mencionado servicio en un entorno con sistema operativo *Linux*:

1. Desde la terminal y situados en el directorio de instalación del software deberemos acceder al directorio *"/bin"*, en el cual entre muchos archivos se encuentra uno cuyo nombre es "*startup.sh*"
2. Se debe ejecutar el mismo, para levantar el servidor local de la máquina, de la siguiente forma: *./startup.sh*
3. Tras ello en cualquier navegador deberemos buscar lo siguiente: *localhost/8080*, tras lo cual se abrirá el *dashboard* del *software*, en el cual se podrán cargar los archivos *XML*, en los cuales se realizarán las consultas pertinentes.

4. Resultados

- Representación porcentual de nuestras categorías con la suma parcial de las mismas, es decir, que porcentaje representa cada categoría con respecto a las demás seleccionadas, para ello se han realizado un total de 4 *XQuery*, una para cada categoría.

```
xquery version "3.1";

let $Businessids1:= doc("Business1.xml")/ystfeed/vespaadd/document/id/text()
let $Businessids2-1:= doc("Business2-1.xml")/ystfeed/vespaadd/document/id/text()
let $Businessids2-2:= doc("Business2-2.xml")/ystfeed/vespaadd/document/id/text()
let $Businessids3:= doc("Business3.xml")/ystfeed/vespaadd/document/id/text()
let $Carsids:= doc("Cars.xml")/ystfeed/vespaadd/document/id/text()
let $Computersids:= doc("Computers.xml")/ystfeed/vespaadd/document/id/text()
let $Entertainmentids:= doc("Entertainment.xml")/ystfeed/vespaadd/document/id/text()

let $total := count ($Businessids1) + count ($Businessids2-1) + count ($Businessids2-2) + count ($Businessids3) + count ($Carsids) + count ($Computersids) + count ($Entertainmentids)
let $businesstotal:= ((count ($Businessids1) + count ($Businessids2-1) + count ($Businessids2-2) + count ($Businessids3)) div $total)

return $businesstotal * 100
```

```
xquery version "3.1";

let $Businessids1:= doc("Business1.xml")/ystfeed/vespaadd/document/id/text()
let $Businessids2-1:= doc("Business2-1.xml")/ystfeed/vespaadd/document/id/text()
let $Businessids2-2:= doc("Business2-2.xml")/ystfeed/vespaadd/document/id/text()
let $Businessids3:= doc("Business3.xml")/ystfeed/vespaadd/document/id/text()
let $Carsids:= doc("Cars.xml")/ystfeed/vespaadd/document/id/text()
let $Computersids:= doc("Computers.xml")/ystfeed/vespaadd/document/id/text()
let $Entertainmentids:= doc("Entertainment.xml")/ystfeed/vespaadd/document/id/text()

let $total := count ($Businessids1) + count ($Businessids2-1) + count ($Businessids2-2) + count ($Businessids3) + count ($Carsids) + count ($Computersids) + count ($Entertainmentids)
let $Carstotal:= (count($Carsids) div $total)

return $Carstotal * 100
```

```
xquery version "3.1";

let $Businessids1:= doc("Business1.xml")/ystfeed/vespaadd/document/id/text()
let $Businessids2-1:= doc("Business2-1.xml")/ystfeed/vespaadd/document/id/text()
let $Businessids2-2:= doc("Business2-2.xml")/ystfeed/vespaadd/document/id/text()
let $Businessids3:= doc("Business3.xml")/ystfeed/vespaadd/document/id/text()
let $Carsids:= doc("Cars.xml")/ystfeed/vespaadd/document/id/text()
let $Computersids:= doc("Computers.xml")/ystfeed/vespaadd/document/id/text()
let $Entertainmentids:= doc("Entertainment.xml")/ystfeed/vespaadd/document/id/text()

let $total := count ($Businessids1) + count ($Businessids2-1) + count ($Businessids2-2) + count ($Businessids3) + count ($Carsids) + count ($Computersids) + count ($Entertainmentids)
let $Computerstotal:= (count($Computersids) div $total)

return $Computerstotal * 100
```

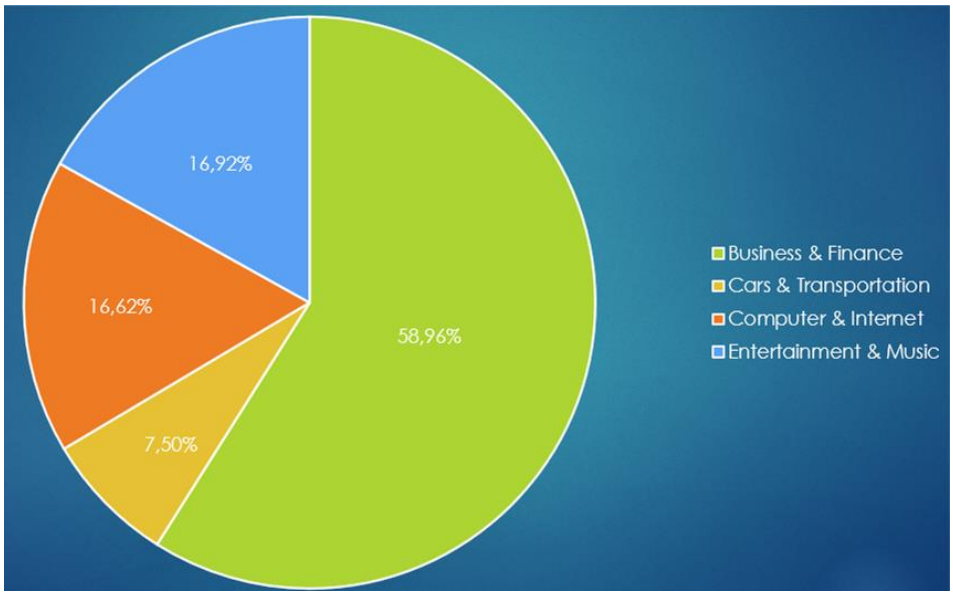
```
xquery version "3.1";

let $Businessids1:= doc("Business1.xml")/ystfeed/vespaadd/document/id/text()
let $Businessids2-1:= doc("Business2-1.xml")/ystfeed/vespaadd/document/id/text()
let $Businessids2-2:= doc("Business2-2.xml")/ystfeed/vespaadd/document/id/text()
let $Businessids3:= doc("Business3.xml")/ystfeed/vespaadd/document/id/text()
let $Carsids:= doc("Cars.xml")/ystfeed/vespaadd/document/id/text()
let $Computersids:= doc("Computers.xml")/ystfeed/vespaadd/document/id/text()
let $Entertainmentids:= doc("Entertainment.xml")/ystfeed/vespaadd/document/id/text()

let $total := count ($Businessids1) + count ($Businessids2-1) + count ($Businessids2-2) + count ($Businessids3) + count ($Carsids) + count ($Computersids) + count ($Entertainmentids)
let $Entertainmenttotal:= (count($Entertainmentids) div $total)

return $Entertainmenttotal * 100
```

Los resultados obtenidos son los siguientes:



- Idiomas en los que se realizaron preguntas en la categoría de *Cars & Transportation*.

```
1  xquery version "3.1";
2
3  let $preguntas := doc("Cars.xml")/ystfeed/vespaadd/document/language/text()
4  for $idiomas in distinct-values($preguntas)
5  return $idiomas
```

Los resultado obtenidos son los siguientes:

```
1 "en-us"
2 "en-uk"
3 "en-ca"
4 "en-au"
5 "en-in"
6 "en-sg"
7 "en-my"
8 "ca-us"
```

- Representación porcentual de preguntas realizadas en inglés de Estados Unidos en la categoría de *Cars & Transportation* y *Entertainment & Music*.

```
xquery version "3.1";

let $ids := doc("Cars.xml")/ystfeed/vespaadd/document/id/text()
let $totalids := count($ids)

let $enus := doc("Cars.xml")/ystfeed/vespaadd/document[language/text()=("en-us")]
let $totalenus := count($enus)


return (($totalenus div $totalids) * 100)
```

```
xquery version "3.1";

let $EntIds := doc("Entertainment.xml")/ystfeed/vespaadd/document/id/text()
let $totalids := count($EntIds)

let $enus:= doc("Entertainment.xml")/ystfeed/vespaadd/document[language/text()=("en-us")]
let $totalenus := count($enus)

return (($totalenus div $totalids) * 100)
```

Cuyos resultados obtenidos en ambos casos son parecidos, superando el 90% de las preguntas de cada categoría.

5. Conclusiones

- Entre las categorías seleccionadas, la que despertaba más interés entre la comunidad era la de *Business & Finance*, representando casi un 60% del total de preguntas seleccionadas entre las 4 categorías.
- En la mayoría de categorías, por lo menos entre las seleccionadas, el idioma predominante para realizar preguntas era el inglés de Estados Unidos, representando aproximadamente el 90%.