Gestão e Tratamento de Informação

Laboratório 12 – Exercícios

Nesta aula de laboratório iremos resolver alguns exercícios relacionados com consolidação de dados semi-estruturados, especificamente abordando o problema da **consolidação de elementos duplicados provenientes de documentos XML**.

Na implementação de um processo de limpeza de dados, a detecção de pares de elementos duplicados é apenas parte do problema, tendo-se que posteriormente se torna necessário (i) gerar grupos de elementos duplicados com base nos pares identificados, e (ii) consolidar os elementos duplicados definidos em cada grupo através de uma regra de agregação definida pelo utilizador que seja adequada ao problema em causa.

Os exercícios desta aula de laboratório serão baseados no mesmo conjunto de dados sobre discos compactos que temos vindo a utilizar nas últimas aulas, tendo-se que iremos também reutilizar as soluções para alguns dos exercícios anteriores.

No URL https://dspace.ist.utl.pt/bitstream/2295/237039/1/cddb_data.xml encontra-se o documento XML contendo informação sobre discos compactos (CDs) e as músicas e artistas que se lhes encontram associados, contendo a descrição de 480 CDs escolhidos aleatoriamente da base de dados FreeDB (http://www.freedb.org/).

No URL https://dspace.ist.utl.pt/bitstream/2295/236816/1/cddb_dups.xml encontra-se o documento XML contendo informação sobre quais as descrições duplicadas de CDs que se encontram no primeiro ficheiro (i.e., os duplicados reais), sejam estes duplicados exactos ou aproximados.

Exercício 1

1.1 - Escreva uma função em XQuery que, usando a similaridade entre CDs tal como calculada em exercícios das aulas de laboratório anteriores (e.g., a similaridade entre pares de CDs com base numa métrica de comparação entre as *strings* dos títulos), permita identificar grupos de CDs muito similares entre si (i.e., grupos de CDs potencialmente duplicados). Na implementação deste exercício, sugere-se uma inspecção ao código da função `local:distinct-nodes()` apresentada na aula de laboratório anterior.

1.2 – Escreva uma expressão XQuery que, com base na função desenvolvida no Exercício 1.1, permita consolidar a informação duplicada que se encontra no documento `cddb_data.xml`. Como resultado deste exercicio, deve ser produzido um documento XML no mesmo formato de representação do documento `cddb_data.xml`, mas em que para cada grupo de CDs duplicados apenas se represente um dos seus CDs.

1.3 – Escreva uma função em XQuery que aceite como entrada uma lista de *strings* e devolva como resultado a maior *string* do conjunto em termos do número de caracteres.

1.4 – Escreva uma função em XQuery que aceite como entrada duas sequencias de elementos XML e que produza como resultado uma sequencia de elementos com a união dos elementos nas duas sequencias de entrada. Na sequencia resultante não devem existir duplicados exactos, e os elementos devem ser apresentados na mesma ordem em que se encontram nas sequencias fornecidas à entrada. Na implementação deste exercício, sugere-se uma inspecção ao código da função `local:union()` apresentada na aula de laboratório anterior.

1.5 - Escreva uma XQuery que permita consolidar a informação duplicada no documento `cddb_data.xml`, consolidando os CDs que sejam duplicados num único registo. Ao consolidar os CDs duplicados, deve procurar obedecer aos seguintes critérios:

- Preencher os vários elementos, combinando a informação dos CDs duplicados.
- Manter o título mais completo (i.e., o título correspondente á maior *string*).
- Usar uma união das listas de faixas em cada CD.

Para resolver este exercício, deve adaptar a sua resolução do Exercício 1.2 e fazer uso das funções XQuery desenvolvidas nos Exercícios 1.3 e 1.4.

1.6 – Altere a função do Exercício 1.1 por forma a que o agrupamento de CDs com base na sua similaridade faça uso do fecho transitivo entre a relação de similaridade entre pares de CDs (e.g., se um CD 1 for similar a um CD 2, e um CD 2 for similar a um CD 3, então devem-se considerar os CDs 1 e 2 e 3 como pertencentes ao mesmo grupo).

1.7 – Altere a expressão XQuery do Exercício 1.5 por forma a considerar o mecanismo de agrupamento de CDs similares desenvolvido no Exercício 1.6.