21103 - Sistemas de Gestão de Bases de Dados
2013-2014
e-fólio C
Resolução e Critérios de Correção

PARA A RESOLUÇÃO DO E-FÓLIO, ACONSELHA-SE QUE LEIA ATENTAMENTE O SEGUINTE:

1) O e-fólio é constituído por 3 perguntas. A cotação global é de 3 valores.

2) O e-fólio deve ser entregue num único ficheiro PDF, não zipado, com fundo branco, com perguntas numeradas e sem necessidade de rodar o texto para o ler. Penalização de 1 a 3 valores.

3) Não são aceites e-fólios manuscritos, i.e. tem penalização de 100%.

4) O nome do ficheiro deve seguir a normal "eFolioC" + <nº estudante> + <nome estudante com o máximo de 3 palavras>. Penalização de 1 a 3 valores.

5) Na primeira página do e-fólio deve constar o nome completo do estudante bem como o seu número. Penalização de 1 a 3 valores.

6) Durante a realização do e-fólio, os estudantes devem concentrar-se na resolução do seu trabalho individual, não sendo permitida a colocação de perguntas ao professor ou entre colegas.

7) A interpretação das perguntas também faz parte da sua resolução, se encontrar alguma ambiguidade deve indicar claramente como foi resolvida.

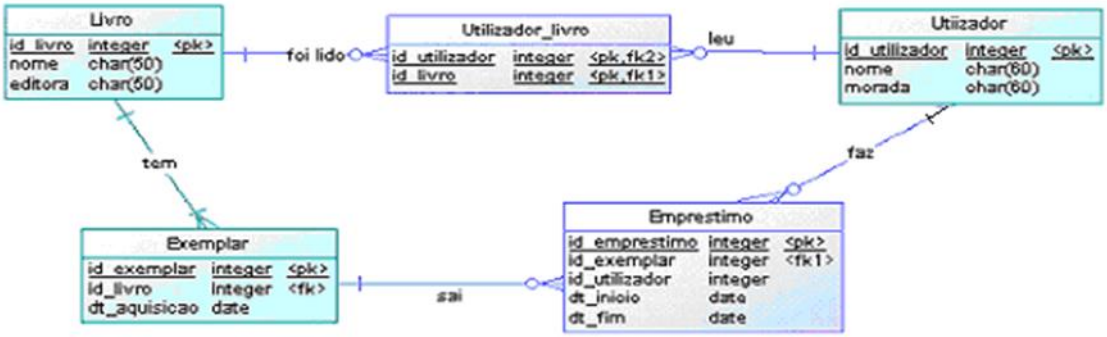
8**)** A legibilidade, a objectividade e a clareza nas respostas serão valorizadas, pelo que, a falta destas qualidades serão penalizadas.

A informação da avaliação do estudante está contida no vetor das cotações:

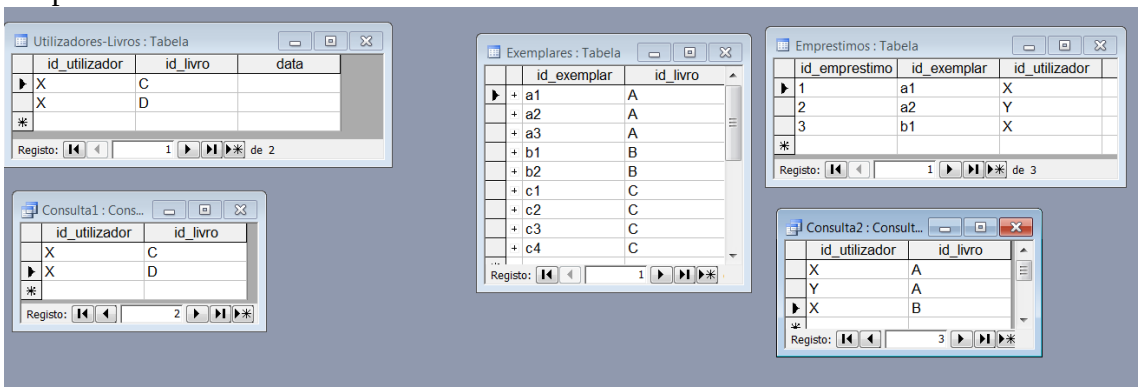
Questão:	1.a	1.b	1.c	2.1	2.2	2.3	3.a	3.b
Cotações:	4	3	3;	3	4	3;	5	5 décimas

1) (1 valor) Na agregação de dados de uma base de dados transacional para um Data Warehouse existe 3 armadilhas no SQL ao utilizar junções (SQL traps):
- junções com múltiplos caminhos
- junções de N:N
- agregação de medidas da tabela pai e da tabela filho

Exemplifique consultas que evidenciem os erros, com dados e resultados, para os seguintes casos:
1.a) junções com múltiplos caminhos



Resposta:



As consultas 1 e 2 devolvem resultados diferentes:
- A consulta 1 utiliza o caminho da tabela Utilizadores-Livros:

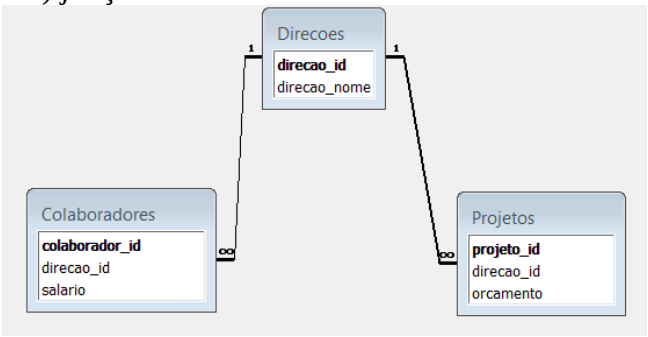
SELECT Utilizadores.id_utilizador, Livros.id_livro
FROM Livros, Utilizadores, [Utilizadores-Livros]
WHERE Utilizadores.id_utilizador = [Utilizadores-Livros].id_utilizador
AND Livros.id_livro = [Utilizadores-Livros].id_livro;

- Enquanto que a tabela 2 utiliza o caminho de Exemplares e Empréstimos:

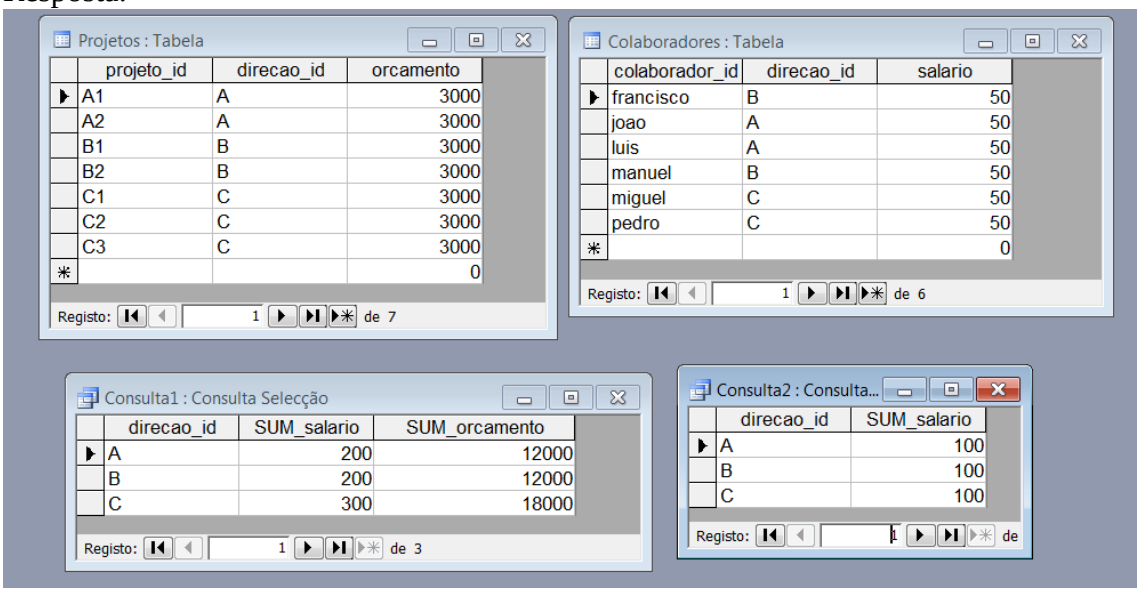
SELECT Utilizadores.id_utilizador, Livros.id_livro
FROM Utilizadores, Livros, Exemplares, Emprestimos
WHERE Livros.id_livro = Exemplares.id_livro
AND Exemplares.id_exemplar = Emprestimos.id_exemplar
AND Utilizadores.id_utilizador = Emprestimos.id_utilizador;

Critérios de correção (4 décimas):
- devem ser apresentadas as consultas, os dados e os resultados errados
- penalização de 1 a 2 décimas se faltarem as consultas, os dados ou os resultados

1.b) junções de N:N



Resposta:



A soma dos salários apresenta valores diferentes nas consultas:
- A consulta 1 apresenta um valor demasiado grande da soma dos salários, depois da junção das 3 tabelas:

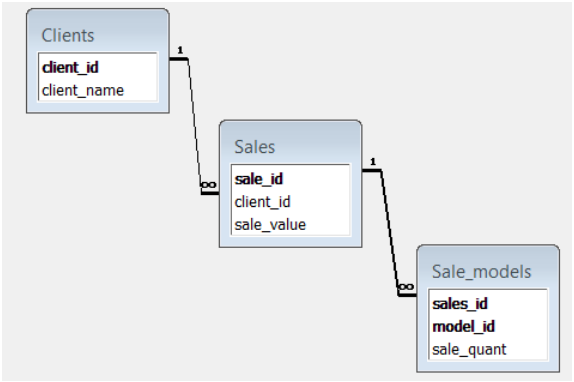
SELECT Direcoes.direcao_id, Sum(Colaboradores.salario) AS SUM_salario, Sum(Projetos.orcamento) AS SUM_orcamento
FROM Direcoes, Colaboradores, Projetos
WHERE Direcoes.direcao_id=Colaboradores.direcao_id
AND Direcoes.direcao_id=Projetos.direcao
GROUP BY Direcoes.direcao_id;

- A consulta 2 apresenta os valores reais:

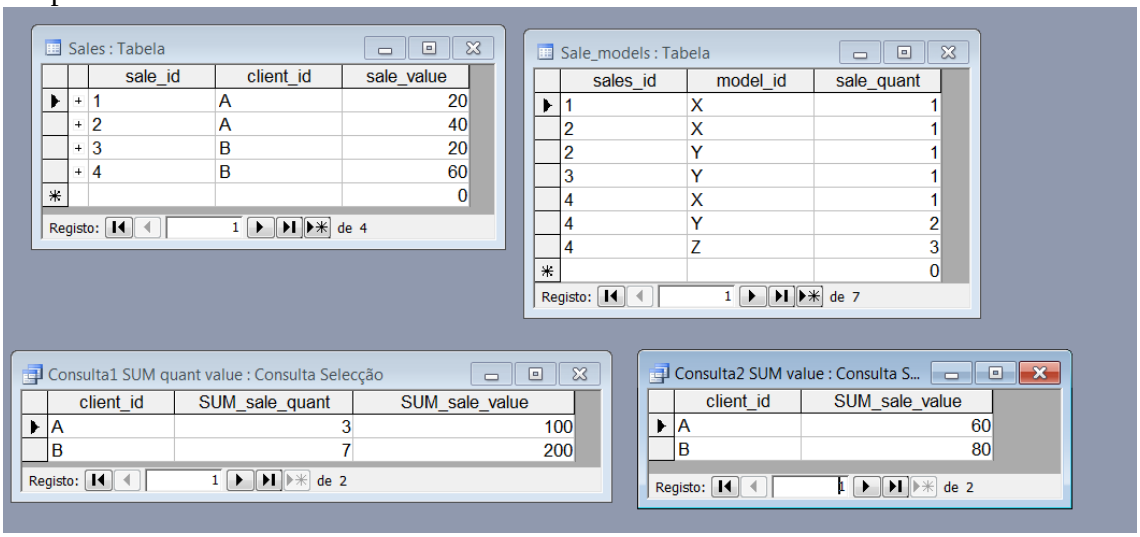
SELECT Direcoes.direcao_id, Sum(Colaboradores.salario) AS SUM_salario
FROM Direcoes, Colaboradores
WHERE Direcoes.direcao_id = Colaboradores.direcao_id
GROUP BY Direcoes.direcao_id;

Critérios de correção (3 décimas):
- devem ser apresentadas as consultas, os dados e os resultados errados
- penalização de 1 a 2 décimas se faltarem as consultas, os dados ou os resultados

1.c) agregação de medidas da tabela pai e da tabela filho



Resposta:



A soma das vendas (sale_value) apresenta valores diferentes nas consultas:
- A consulta 1 apresenta um valor demasiado grande da soma das vendas, depois da junção das 3 tabelas:

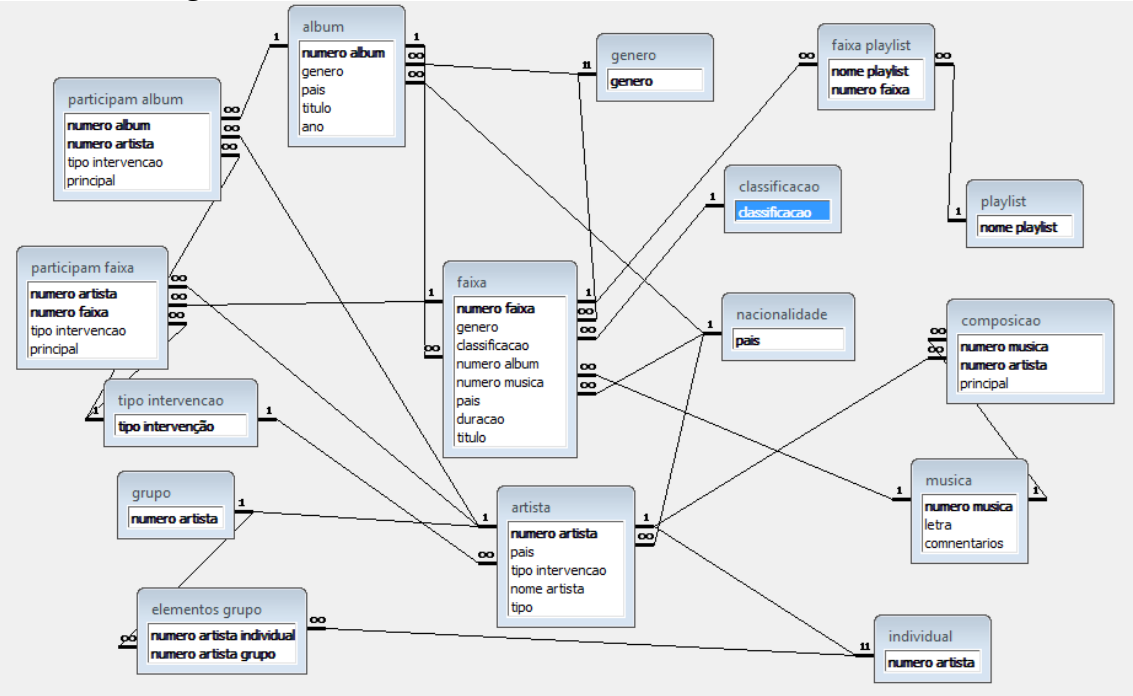
SELECT Clients.client_id, Sum(Sale_models.sale_quant) AS SUM_sale_quant, Sum(Sales.sale_value) AS SUM_sale_value
FROM Clients, Sales, Sale_models
WHERE Clients.client_id = Sales.client_id
AND Sales.sale_id = Sale_models.sales_id
GROUP BY Clients.client_id;

- A consulta 2 apresenta os valores reais:

SELECT Clients.client_id, Sum(Sales.sale_value) AS SUM_sale_value
FROM Clients, Sales
WHERE Clients.client_id = Sales.client_id
GROUP BY Clients.client_id;

Critérios de correção (3 décimas):
- devem ser apresentadas as consultas, os dados e os resultados errados
- penalização de 1 a 2 décimas se faltarem as consultas, os dados ou os resultados

2) (1 valor) *Data Warehousing*
Considere a seguinte bases de dados relativa a músicas num leitor de MP3:

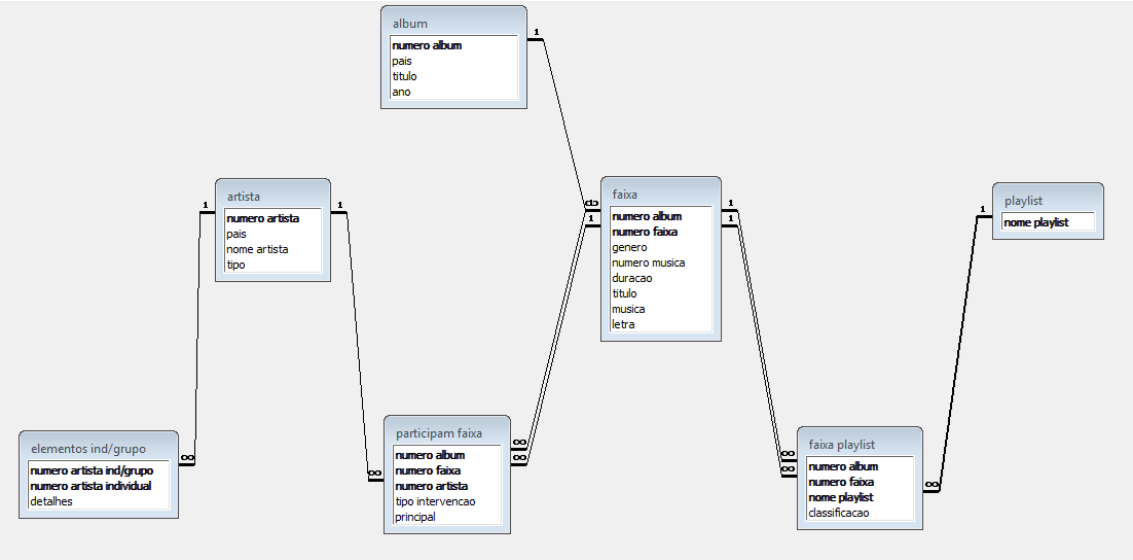


2.1- Desenhe uma base de dados transacional equivalente, na 3ª forma normal. De seguida remova a eventual transitividade que exista no esquema base de dados. Faça o carregamento de dados. Na representação gráfica das ligações de 1:N, a tabela com uma única linha é desenhada em cima e a tabela com várias linhas é desenhada por baixo.

Resposta:

Para criar um esquema sem múltiplos caminhos foram aplicados os seguintes passos:
i) dado o ciclo <artista, grupo, individual, elementos_grupo> foram removidas as relações 1:1; o resultado resume-se a duas tabelas com uma ligação artista-elementos_grupo (individual ou grupo);
ii) dado ciclo <álbum, faixa, género> retirar a ligação género-álbum já que existe uma ligação género-faixa
iii) dado o ciclo <tipo_intervenção, participam_faixa, artista> retirar a ligação tipo_intervencao - artista, já que existe um tipo de intervenção na participam na faixa;
iv) a tabela país tem várias ligações; foram criados dois aliás para as tabelas álbum e faixa, a tabela original fica ligada a artista;
v) remover a tabela participam_album já que existe a mesma informação na tabela participam_faixa;
vi) remover a tabela composição podendo esta informação ser atualizada em participam_faixa.tipo_intervenção.

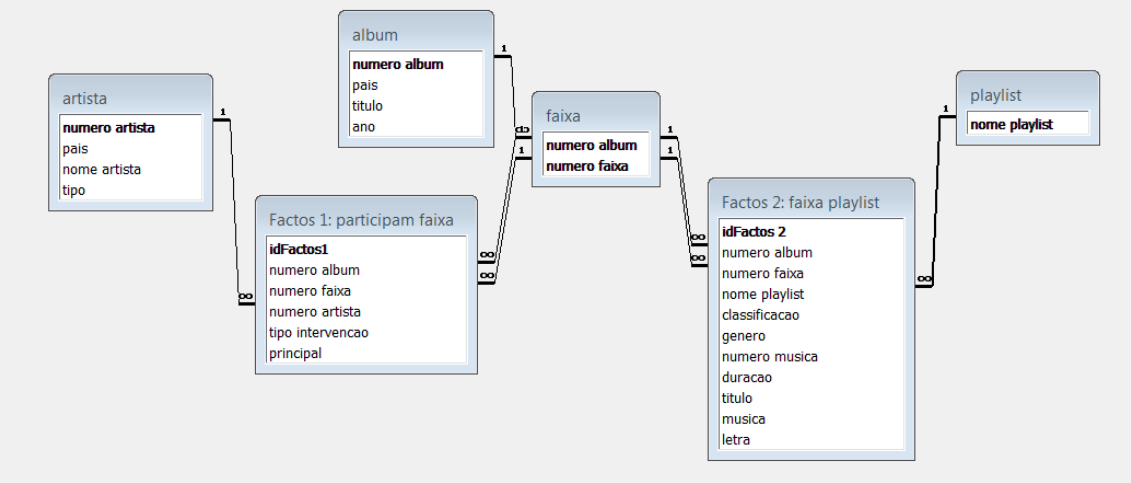
Obtemos finalmente a seguinte base de dados:



Critérios de correção (3 décimas):
- devem ser apresentada a base de dados sem caminhos múltiplos: 3 décimas
- penalização de 1 a 2 décimas para caminhos múltiplos

2.2- Pretendemos desenhar um “Data Warehouse” relacional em estrela ou em constelação, i.e. com duas ou mais estrelas. Defina a(s) tabela(s) de factos e mostre a tabela depois da desnormalização dos dados. Defina as dimensões com os níveis de agregação para o “Data Warehouse” relacional. Apresente a(s) tabela(s) de factos associada às dimensões. Ao juntar as tabelas transacionais tenha em consideração as eventuais armadilhas referidas na pergunta anterior.

Resposta:

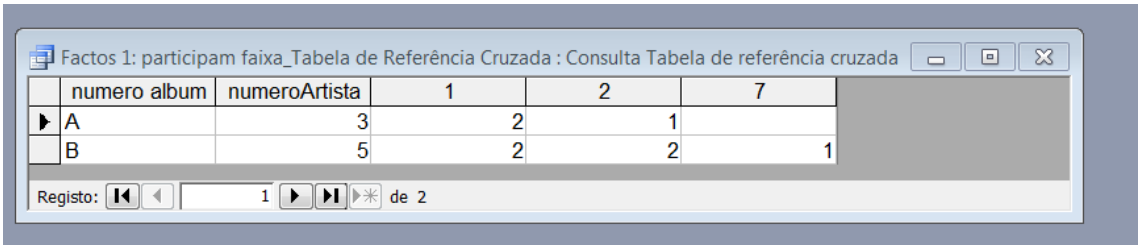


Critérios de correção (4 décimas):
- devem ser apresentada o Data Warehouse com 2 tabelas de factos: 4 décimas
- DW com 1 tabela factos: 2 décimas

2.3- Crie duas perguntas e traduza para SQL utilizando pelo menos duas dimensões.

Resposta:

a) Frequência dos artistas nos álbuns

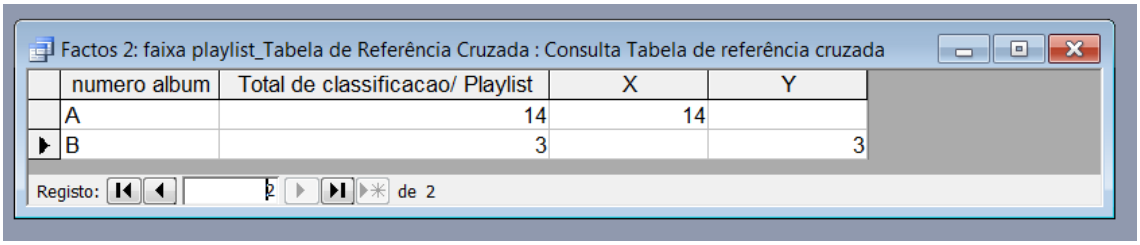

Factos 1: participam faixa_Tabela de Referência Cruzada : Consulta Tabela de referência cruzada

numero album	numeroArtista	1	2	7
A	3	2	1	
B	5	2	2	1

Registo: 1 de 2

TRANSFORM Count([Factos 1: participam faixa].idFactos1)
SELECT [Factos 1: participam faixa].[numero album], Count([Factos 1: participam faixa].idFactos1) AS numeroArtista
FROM [Factos 1: participam faixa]
GROUP BY [Factos 1: participam faixa].[numero album]
PIVOT [Factos 1: participam faixa].[numero artista];

b) Soma das classificações por álbum e playlist


Factos 2: faixa playlist_Tabela de Referência Cruzada : Consulta Tabela de referência cruzada

numero album	Total de classificacao/ Playlist	X	Y
A	14	14	
B	3		3

Registo: 2 de 2

TRANSFORM Sum([Factos 2: faixa playlist].classificacao) AS SomaDeclassificacao
SELECT [Factos 2: faixa playlist].[numero album], SUM([Factos 2: faixa playlist].classificacao) AS [Total de classificacao/ Playlist]
FROM [Factos 2: faixa playlist]
GROUP BY [Factos 2: faixa playlist].[numero album]
PIVOT [Factos 2: faixa playlist].[nome playlist];

Critérios de correção (3 décimas):
- devem ser apresentados os resultados e as consultas SQL utilizador o Pivot

3) (1 valor) *Information Retrieval*

3.a) Num "corpus" com 1.000 documentos, para a consulta Q="curso e-learning", qual a ordem decrescente de relevância os seguintes documentos?

Doc1 = "O ensino através do e-learning pode ser síncrono ou assíncrono."
Doc2 = "Os cursos de e-learning tornam possível a cobertura de públicos geograficamente dispersos."
Doc3 = "São feitas críticas aos cursos de e-learning devido à necessidade de maior disciplina e auto-organização por parte do estudante."

Resposta:

Seja n(d) o número de termos num documento "d" e n(d,t) o número de termos "t" num documento "d", em que a Frequência de um Termo "t" num documento "d" é dado no manual por:

Seja n(t) o número de documentos que contêm o termo "t" e N o número total de documentos, onde o Inverse Document Frequency (IDF) de Salton & Buckley 1988 é dado por:

A relevância de um termo "t" num documento "d" é dado por:

$$\text{TF-IDF}(d,t) = \text{TF}(d,t).\text{IDF}(t)$$

Para a resolução deste problema é necessário remover o plural (ou o singular) dos termos e ignorar a lista de palavras conhecida por "stop words".

stop terms
a
à
aos
as
da
de
de
do
e
o
os
ou

A relevância "r" de um documento "d" para um conjunto de termos Q é dado por:

$$r(d,Q) = \sum_{t \in Q} TF(d,t) \times IDF(t)$$

Em primeiro lugar, vamos calcular os valores de IDF:

		curso	e-learning
# documentos com termo t	n(t)	2	3
# documentos	N	1.000	1.000
	IDF(t)=log(N/n(t), 10)	2,70	2,52

De seguida cálculos a relevância dos documentos: r(Doc1,Q), r(Doc2,Q) e r(Doc3,Q), utilizado a expressão: $r(d,Q) = \sum_{t \in Q} TF(d,t) \times IDF(t)$

Doc1		curso	e-learning	r(Doc1, Q)
# terms in the document d	n(d)	7	7	
# term t in document d	n(d,t)	0	1	
	TF(d,t)=ln(1+n(d,t)/n(d))	0,00	0,13	
	TF-IDF(d,t)	0,00	0,34	0,34

Doc2		curso	e-learning	r(Doc1, Q)
# terms in the document d	n(d)	8	8	
# term t in document d	n(d,t)	1	1	
	TF(d,t)=ln(1+n(d,t)/n(d))	0,12	0,12	
	TF-IDF(d,t)	0,32	0,30	0,62

Doc3		curso	e-learning	r(Doc1, Q)
# terms in the document d	n(d)	12	12	
# term t in document d	n(d,t)	1	1	
	TF(d,t)=ln(1+n(d,t)/n(d))	0,08	0,08	
	TF-IDF(d,t)	0,22	0,20	0,42

Obtemos assim a seguinte ordem decrescente de relevância: Doc2, Doc3, Doc1

Critérios de Correção (5 décimas):
- pretende-se que o estudante utilize a expressão r(d,Q) = Σ TF(d,t) . IDF(t)
- somatório (1 décimas); TF (2 décimas); IDF (2 décimas)
- penalização 1 a 2 décimas para cálculos errados
- penalização 1 a 2 décimas para cálculos desnecessários

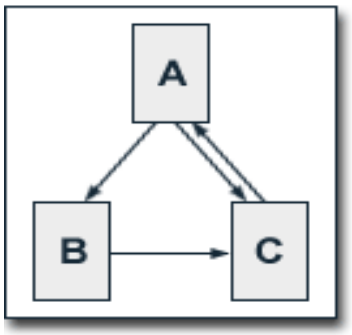
3.b) Descreva o que entende por PageRank.

O algoritmo original de PageRank descrito por Lawrence Page and Sergey Brin em 1995 é dado por:

PR(A) = (1-d) + d (PR(T1)/C(T1) + ... + PR(Tn)/C(Tn))

onde
- PR(A) é o PageRank da página A,
- PR(Ti) é o PageRank das páginas Ti que estão ligadas (apontam) para a página A,
- C(Ti) é o número de apontadores ("outbound links") na página Ti
- d é o fator de amortecimento que varia em 0 e 1.

Exemplo:



Seja d=0.5,

PR(A) = 0.5 + 0.5 (PR(C) /1)
PR(B) = 0.5 + 0.5 (PR(A) / 2)
PR(C) = 0.5 + 0.5 (PR(A) / 2 + PR(B) /1)

Resolvendo o sistema de 3 equações e 3 incógnitas obtemos os seguintes PR:

PR(A) = 14/13 = 1.07692308
PR(B) = 10/13 = 0.76923077
PR(C) = 15/13 = 1.15384615

Fonte: http://pr.efactory.de/e-pagerank-algorithm.shtml

Critérios de Correção (5 décimas):
- pretende-se vá para além da explicação do manual e que exemplifique o algoritmo;
- explicação geral 3 décimas
- exemplo 2 décimas