UNIVERSIDADE ABERTA
www.uab.pt

e-Fólio B

U.C. 21103

Sistemas de Gestão de Bases de Dados

2022-2023

Resolução e Critérios de Correção

INSTRUÇÕES

1) O e-fólio é constituído por 5 perguntas. A cotação global é de 5 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. Cada pergunta com uma ou mais páginas, deve ser iniciada numa nova página. Penalização de 10% a 100%.

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

4) O nome do ficheiro deve seguir a normal "eFolioB" + <nº estudante> + <nome estudante com o máximo de 3 palavras>. Penalização de 10% a 100%.

5) Na primeira página do e-fólio deve constar o nome completo do estudante bem como o seu número. Penalização de 10% a 100%.

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.

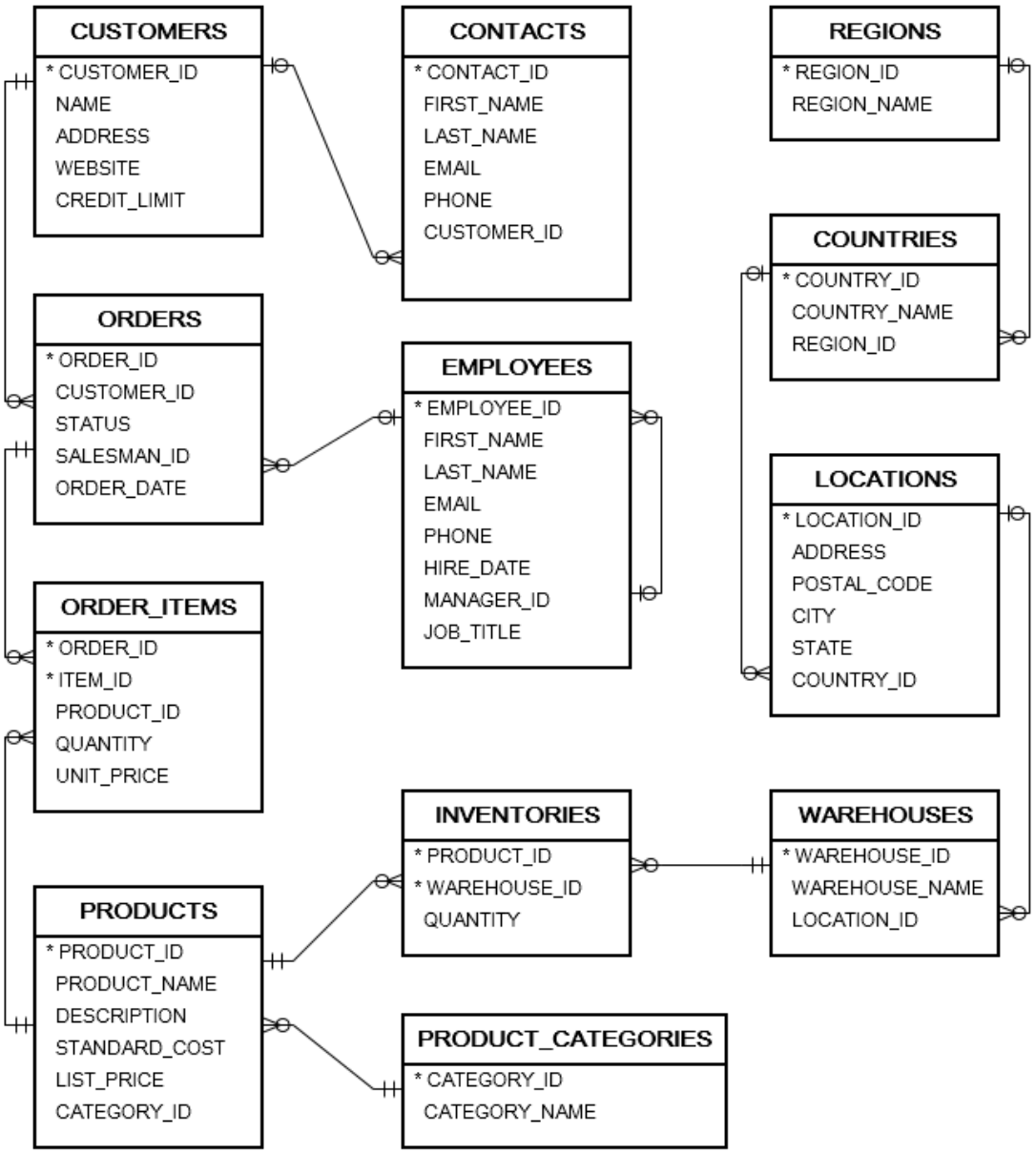
8) A legibilidade, a objetividade e a clareza nas respostas serão valorizadas, pelo que, a falta destas qualidades será penalizada.

9) Critérios de correção gerais: todas as respostas devem ser justificadas, incluir imagens e exemplos com vista a clarificar os argumentos expostos.

Vetor Cotações

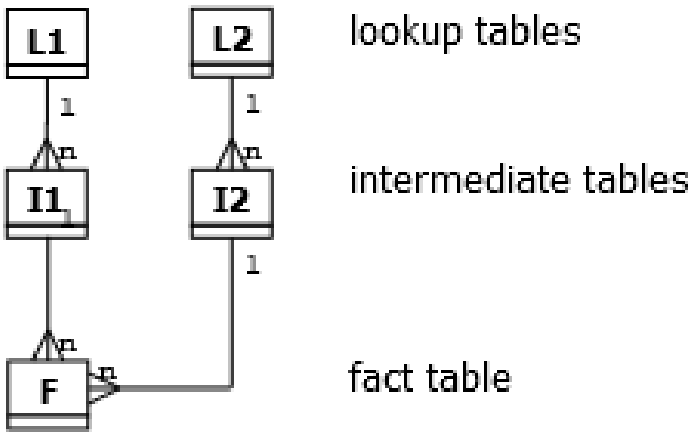
1	2	3,	4	5	pergunta
10	10	10,	10	10	décimas

Considere a seguinte base de dados de uma empresa de distribuição de produtos, para as perguntas seguintes:



1) (1 valor) Desnormalização
Considere a base de dados de uma empresa de distribuição de produtos.

1.a) Represente graficamente as ligações de 1:N, a tabela com uma única linha é desenhada em cima e a tabela com várias linhas é desenhada por baixo. Depois de representar as tabelas classifique-as segundo a tipologia indicada.



1.b). Encontre a 1FD (1ª forma desnormalizada) e a 2FD (2ª forma desnormalizada).
1FD – obter uma poli-árvore, replicando as tabelas que forem necessárias;
2FD – obter árvores separadas, replicando as tabelas que forem necessárias.
Justifique a resposta.

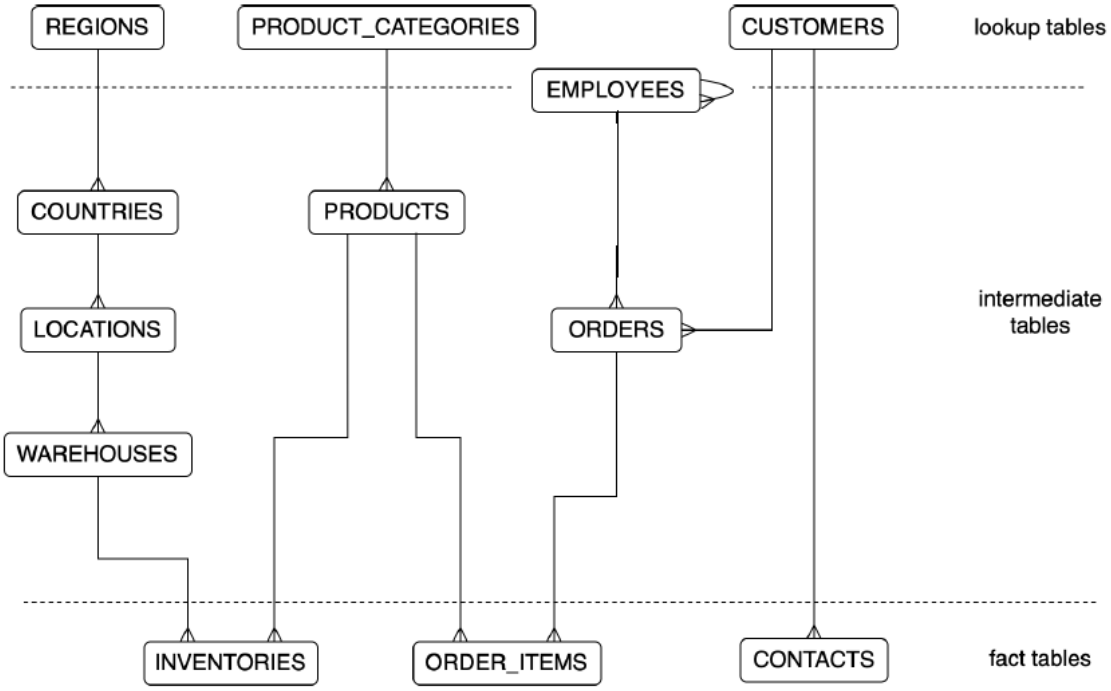
1.c) Considerando que:

- Aditivos: são atributos que podem ser agregados (somados) por todas as dimensões, ex: valor da venda (usar Sum() sempre)
- Semi-aditivos: são atributos que podem ser agregados (somados) por algumas as dimensões, ex: quantidade (usar Sum() em condições particulares)
- Não-aditivos: são atributos que não podem ser agregados (somados), ex: preço unitário (usar Average() por exemplo)
- Sem factos: só existem identificadores (usar a função Count() dos identificadores).

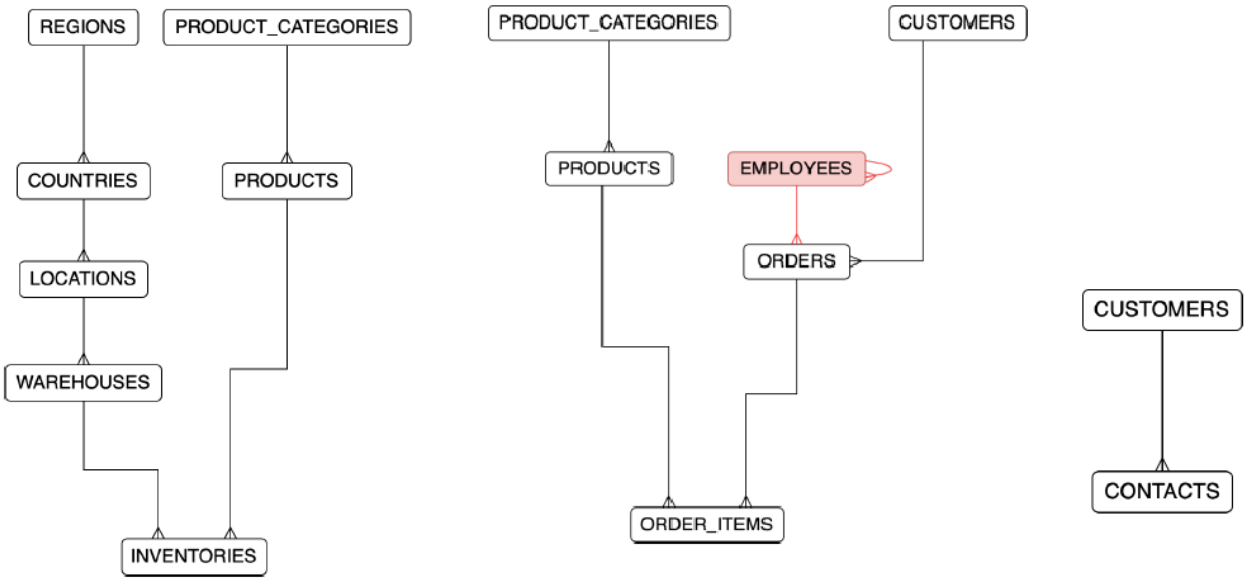
Para as tabelas de factos encontradas defina os atributos aditivos, semi-aditivos, não-aditivos e sem factos. Justifique a resposta.

Resposta:

1.a) Tipologia



1.b) 2FD (2ª forma desnormalizada)



1.c) Tipos atributos:

A tabela INVENTORIES tem um atributo **semi-aditivo**, a Quantidade, todos os outros atributos são sem factos.

A tabela ORDER_ITEMS: o produto de Quantidade*Preço obtemos o Valor da Encomenda, tem atributos **aditivos**.

A tabela CONTACTS só tem atributos **sem factos** já que só existem identificadores (só se pode utilizar a função Count dos identificadores).

Critérios de correção:
- a) 4 décimas, base dados na 3FN e tipologia das tabelas
- b) 3 décimas, encontrar a 2 DF
- c) 3 décima, tipo de atributos
- erros, omissões, redundâncias ou apresentação desadequada: -20% a -100%

2) (1 valor) Data Warehouse
Considere a base de dados de uma empresa de distribuição de produtos.

2.a) Pretendemos desenhar um "Data Warehouse" relacional em estrela ou em constelação, i.e. com duas ou mais estrelas com a maior granularidade possível.

Preencha a 'bus matrix' (ou 'business matrix').

		dimension 1	dimension 2	dimension 3	dimension 4	dimension 5	dimension 6
database	fact table						

Defina a(s) tabela(s) de factos;
Defina as dimensões;
Apresente a(s) tabela(s) de factos associada às dimensões.

2.b) Formule duas perguntas em português corrente que utilize pelo menos duas dimensões do "Data Warehouse".

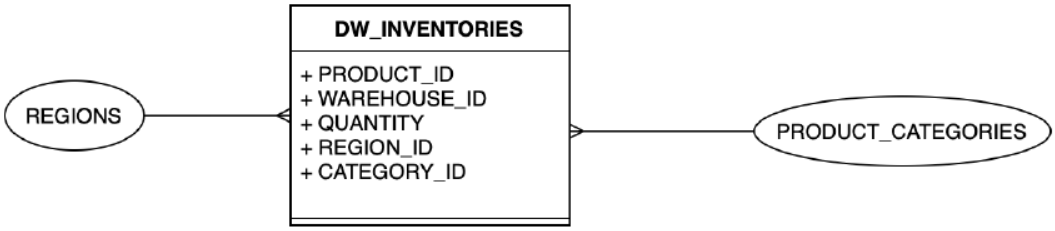
2.c) Traduza as perguntas para SQL utilizando a sintaxe "TRANSFORM … SELECT … PIVOT …" e apresente os resultados da consulta.

Resposta:

2.a) Bus matrix

	Dimensão 1	Dimensão 2	Dimensão 3
Tabela de Factos	REGIONS	PRODUCT_CATEGORIES	CUSTOMERS
INVENTORIES	X	X	
ORDER_ITEMS		X	X
CONTACTS			X

Em relação à tabela de factos INVENTORIES temos:



2.b) Perguntas português

Pergunta 1 - Qual é o número total de produtos em inventário por região e por categoria de produto?

2.c) Perguntas SQL com resultado da consulta

```
TRANSFORM Sum(DW_inventories.quantity) AS SumOfquantity
SELECT DW_inventories.region_id, Sum(DW_inventories.quantity) AS [TQuantity]
FROM DW_inventories
GROUP BY DW_inventories.region_id
PIVOT DW_inventories.category_id;
```

DW_inventories_Crosstab

region_id	Total Of qua	X	Y
A	5813	3241	2572
B	15467	8233	7234

Critérios de correção:
- a) 4 décimas, bus matrix
- b) 2 décimas, 2 perguntas português
- c) 4 décimas, 2 perguntas SQL com resultados
- erros, omissões, redundâncias ou apresentação desadequada: -20% a -100%

3) (1 valor) Leitura de bases de dados
Considere o recurso educativo "Como ler (e interpretar) uma base de dados de grandes dimensões" em https://repositorioaberto.uab.pt/handle/10400.2/12141.

3.a) O que entende por tríade?

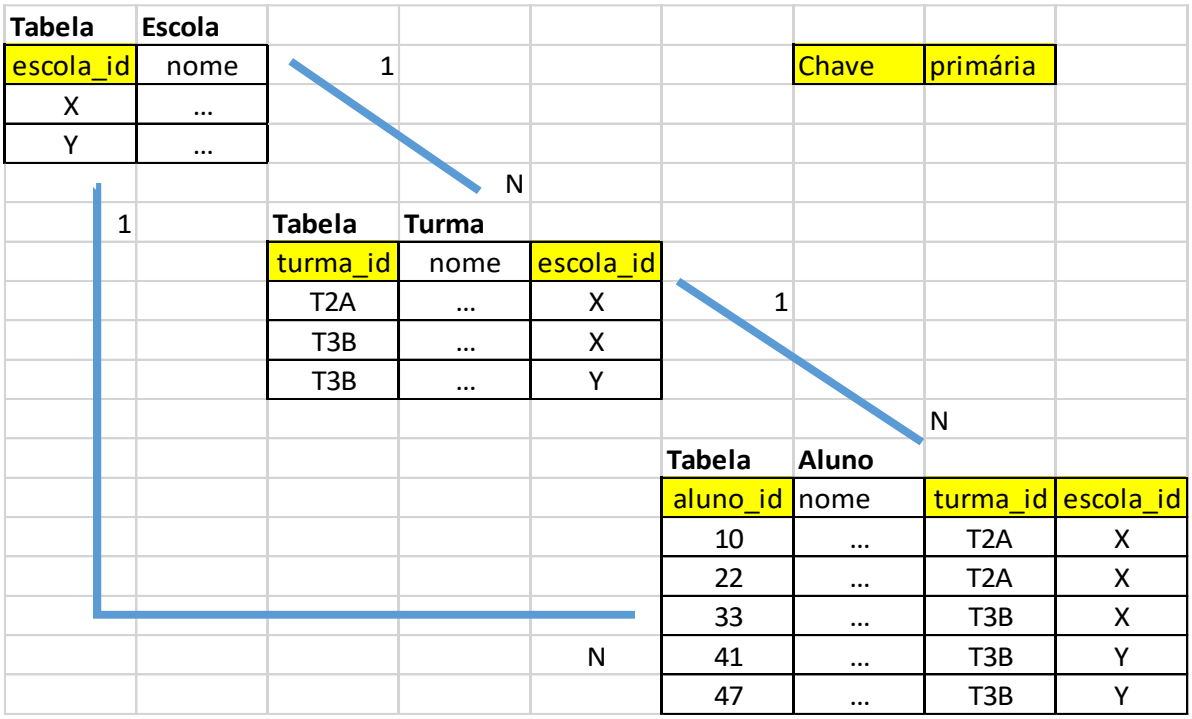
3.b) Exemplifique uma tríade redundante, onde cada tabela deve ter pelo menos 5 registos.

3.c) Encontre de seguida a solução não redundante.

Resposta:

3.a) A tríade é um caso particular do MAPP (problema de acesso por múltiplos caminhos), onde num conjunto de 3 tabelas com 3 relações consegue-se chegar da primeira à última tabela por dois caminhos.

3.b)



3.c)

Tabela	Escola
escola_id	nome
X	...
Y	...

1 — N

Tabela	Turma	
turma_id	nome	escola_id
T2A	...	X
T3B	...	X
T3B	...	Y

1 — N

Tabela	Aluno	
aluno_id	nome	turma_id
10	...	T2A
22	...	T2A
33	...	T3B
41	...	T3B
47	...	T3B

Critérios de correção:
- a) 3 décimas, definição
- b) 3 décimas, 2 solução redundante
- c) 4 décimas, 2 solução não redudante
- erros, omissões, redundâncias ou apresentação desadequada: -20% a -100%

4) (1 valor) Data Mining

4.a) O que entende por técnicas supervisionadas e não-supervisionadas em Data Mining?

4.b) Com base nas "Lecture Notes: Ciências dos Dados", formule duas questões. Uma para base de dados e outra para data mining no exemplo da empresa de distribuição de produtos.

Respostas parcial:

4.a) As técnicas supervisionadas e não supervisionadas são tipos de algoritmos utilizados em 'data mining'.

Os algoritmos supervisionados são os considerados algoritmos preditivos, ou seja, têm como objetivo prever eventos futuros. Estes algoritmos carecem de atributos discriminantes. Podemos referir os algoritmos de Classificação e de Regressão:

- Algoritmos de classificação …
- Algoritmos de regressão …

Os algoritmos não supervisionados são considerados algoritmos descritivos, ou seja, pretendem descrever eventos passados. Estes algoritmos não são orientados a qualquer atributo em especial. Podemos referir os algoritmos de Segmentação e Associação:

- Algoritmos de Segmentação …
- Algoritmos de Associação …

4.b) Perguntas

Pergunta de Base de dados: Quais os 10 produtos mais vendidos?

Pergunta de Data Mining:
Quais os 3 produtos mais vendidos em conjunto? (Associação)
Quais os grupos dos clientes que existem? (Clustering)

Critérios de correção:
- a) 6 décimas, definição
- b) 4 décimas, 2 perguntas
- erros, omissões, redundâncias ou apresentação desadequada: -20% a -100%

5) (1 valor) Information Retrieval

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

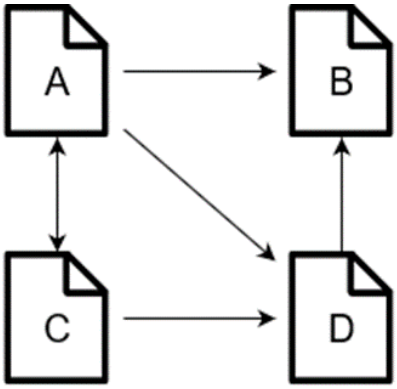
$$P[j] = \delta + (1 - \delta) * \sum_{i=1}^{N} (T[i,j] * P[i])$$

com δ = 0,5.

Encontre a ordenação das páginas para o seguinte conjunto de dados:

T[i,j]

	A	B	C	D	soma
A	0	1/3	1/3	1/3	1
B	0	0	0	0	0
C	1/2	0	0	1/2	1
D	0	1	0	0	1



Resposta:

Para a página A temos:

$$P[A] = \frac{1}{2} + \frac{1}{2} \cdot \sum_{i=A}^{D} (T[i,A] \cdot P[i]) \Leftrightarrow$$

$$\Leftrightarrow P[A] = \frac{1}{2} + \frac{1}{2} (T[A,A] \cdot P[A] + T[B,A] \cdot P[B] + T[C,A] \cdot P[C] + T[D,A] \cdot P[D]) \Leftrightarrow$$

$$\Leftrightarrow P[A] = \frac{1}{2} + \frac{1}{2} \left(0 \cdot P[A] + 0 \cdot P[B] + \frac{1}{2} \cdot P[C] + 0 \cdot P[D]\right) \Leftrightarrow$$

$$\Leftrightarrow P[A] = \frac{1}{2} + \frac{1}{2} \left(\frac{1}{2} \cdot P[C]\right) \Leftrightarrow$$

$$\Leftrightarrow P[A] = \frac{1}{2} + \frac{1}{4} \cdot P[C]$$

Recorrendo à expressão:

$$P[j] = \delta + (1 - \delta) * \sum_{i=1}^{N} (T[i,j] * P[i])$$

iteração	A	B	C	D
0	1	1	1	1
1	0.75	1.166667	0.666667	0.916667
2	0.666667	1.083333	0.625	0.791667
3	0.65625	1.006944	0.611111	0.767361
4	0.652778	0.993056	0.609375	0.762153
5	0.652344	0.989873	0.608796	0.76114
6	0.652199	0.989294	0.608724	0.760923
7	0.652181	0.989161	0.6087	0.760881
8	0.652175	0.989137	0.608697	0.760872
9	0.652174	0.989132	0.608696	0.76087
10	0.652174	0.989131	0.608696	0.76087
11	0.652174	0.98913	0.608696	0.76087
12	0.652174	0.98913	0.608696	0.76087
13	0.652174	0.98913	0.608696	0.76087
14	0.652174	0.98913	0.608696	0.76087
15	0.652174	0.98913	0.608696	0.76087
16	0.652174	0.98913	0.608696	0.76087
17	0.652174	0.98913	0.608696	0.76087
18	0.652174	0.98913	0.608696	0.76087
19	0.652174	0.98913	0.608696	0.76087
20	0.652174	0.98913	0.608696	0.76087
21	0.652174	0.98913	0.608696	0.76087
22	0.652174	0.98913	0.608696	0.76087
23	0.652174	0.98913	0.608696	0.76087
PageRank	3	1	4	2

Solução: B > D > A > C

Critérios de correção:
- a) 5 décimas, fórmulas
- b) 5 décimas, resultado
- erros, omissões, redundâncias ou apresentação desadequada: -20% a -100%