21103 - Sistemas de Gestão de Bases de Dados
2015-2016
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:

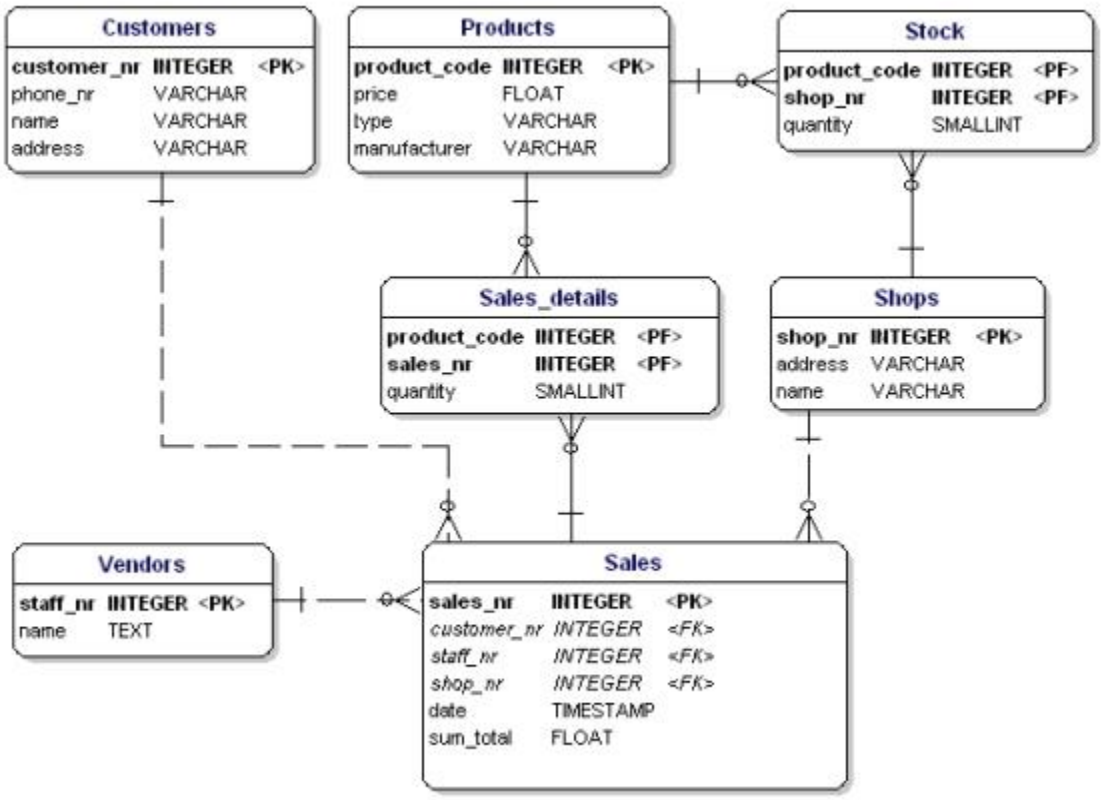
Questão:	1.a	1.b	2.a,b,c	2.d,e	3
Cotações:	5	5,	6	4,	10 décimas

1) (1 valor) No processo de extração de dados de uma base de dados transacional existem 2 tipos de armadilhas no SQL (SQL traps):

- junções com múltiplos caminhos (“multiple access path problem”, “loop”)

- junções com agregações de dados de 2 tabelas (“connection trap”)

Dada a seguinte base de dados transacional de vendas de produtos

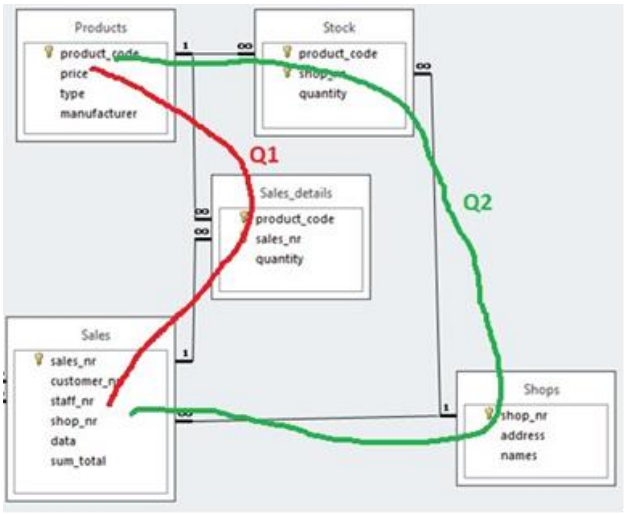


1.a) Para a base de dados da figura exemplifique uma consulta que evidencie a armadilha de junções com múltiplos caminhos, com dados e resultados.

Resp:

Nas respostas às consultas pretende-se saber em que vendas estiveram envolvidos os produtos:

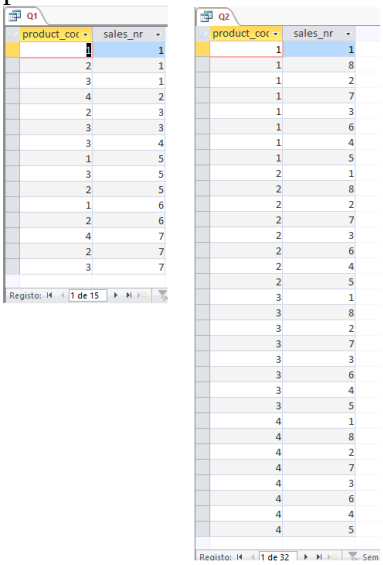
Q1: $\prod_{product\_code,sales\_nr}(Product \mid\!><\!\mid Sales\_details \mid\!><\!\mid Sales)$
Q2: $\prod_{productcode,salesnr}(Product \mid\!><\!\mid stocks \mid\!><\!\mid Shops \mid\!><\!\mid Sales)$



EM SQL teremos para Q1:
SELECT P.product_code, S.sales_nr
FROM Products P, Sales_details SD, Sales S
WHERE P.product_code = SD.product_code
AND SD.sales_nr = S.sales_nr;

E para Q2:
SELECT P.product_code, S.sales_nr
FROM Products P, Stock ST, Shops SH, Sales S
WHERE P.product_code=ST.product_code
AND ST.shop_nr=SH.shop_nr
AND SH.shop_nr=S.shop_nr;

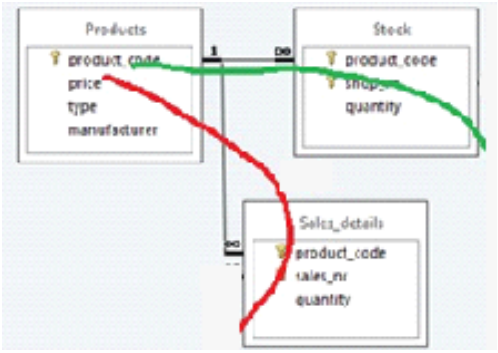
Os resultados para Q1 e Q2 para os dados carregados demonstram a inconsistência nas respostas, evidenciado que para consultas com caminhos de junções diferentes é possível obter resultados diferentes.



Critérios de correção 1.a (5 décimas):
- devem ser apresentadas as consultas e exemplo de resultados errados
- penalização de 2 a 3 décimas se faltarem as consultas ou os resultados

1.b) Para a base de dados da figura exemplifique uma consulta que evidencie a junção com agregações de dados de 2 tabelas, com dados e resultados.

Resp:



Para esta demonstração serão utilizados os dados previamente carregados e as tabelas: Products, Stock e Sales_details.

Executando as seguintes consultas, cada uma envolvendo apenas 2 das 3 tabelas, o resultado é correto:

Consulta 1:

```
SELECT P.product_code, SUM(SD.quantity) AS C1
FROM Products  P, Sales_details  SD
WHERE P.product_code=SD.product_code
GROUP BY P.product_code;
```

bC1

product_coc	C1
1	9
2	9
3	15
4	5

Consulta 2:

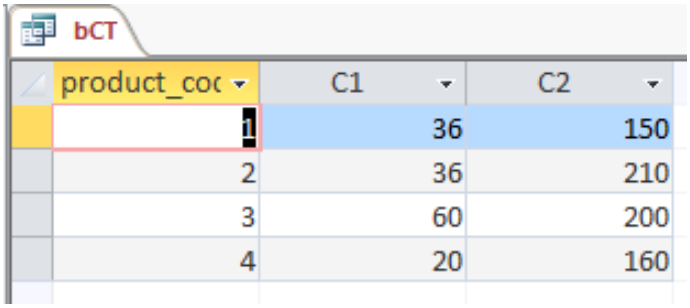
```
SELECT P.product_code, SUM(S.quantity) AS C2
FROM Products P, Stock  S
WHERE P.product_code=S.product_code
GROUP BY P.product_code;
```

bC2

product_coc	C2
1	50
2	42
3	40
4	80

No entanto, quando se juntam as três tabelas, o resultado já não está correto:

```
SELECT P.product_code, SUM(SD.quantity) AS C1, SUM(S.quantity) AS C2 FROM
Products AS P, Sales_details AS SD, Stock AS S
WHERE P.product_code=SD.product_code
AND P.product_code=S.product_code
GROUP BY P.product_code;
```



bCT

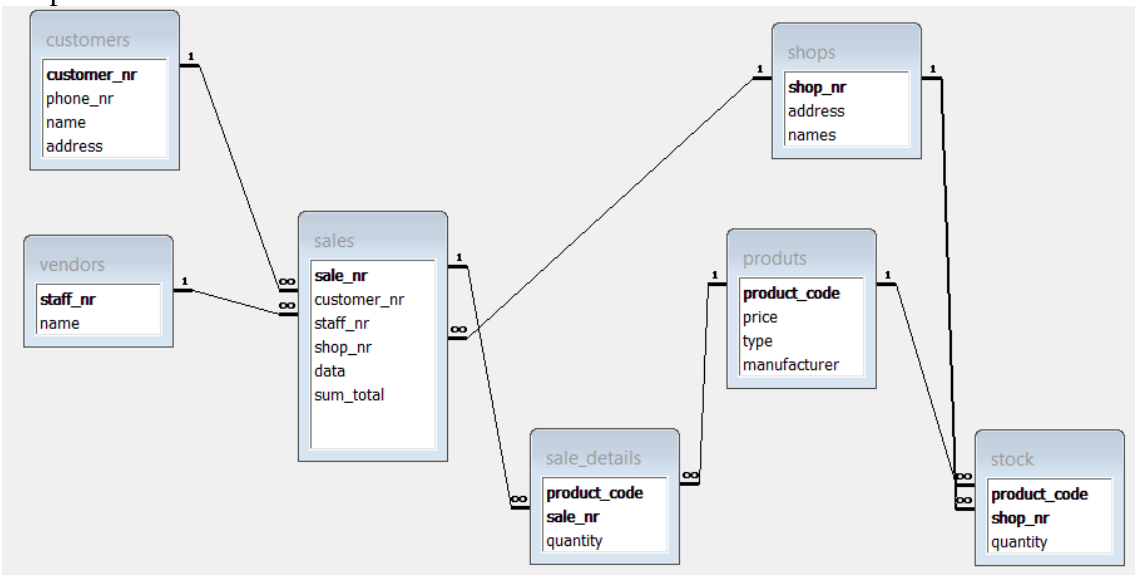
product_coc	C1	C2
1	36	150
2	36	210
3	60	200
4	20	160

Critérios de correção 1.b (5 décimas):
- devem ser apresentadas as consultas e os resultados errados
- penalização de 2 a 3 décimas se faltarem as consultas ou os resultados

2) (1 valor) Para a mesma base de dados transacional de vendas de produtos:

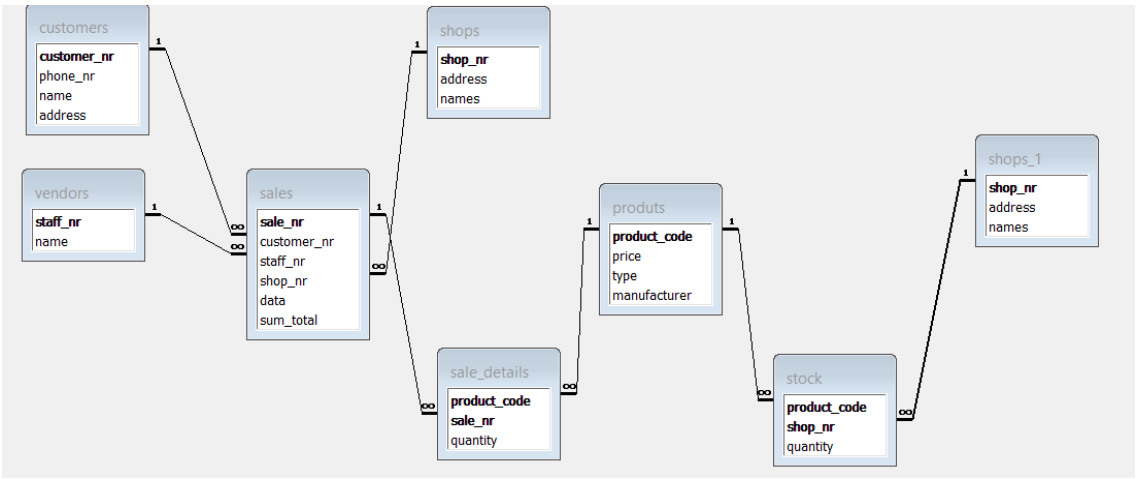
2.a) Desenhe uma base de dados transacional equivalente, na 3ª forma normal. Faça o carregamento de dados. Na representação gráfica das ligações de 1:N, a tabela com uma única linha é desenhada em cima e a tabela com várias linhas é desenhada por baixo.

Resp:



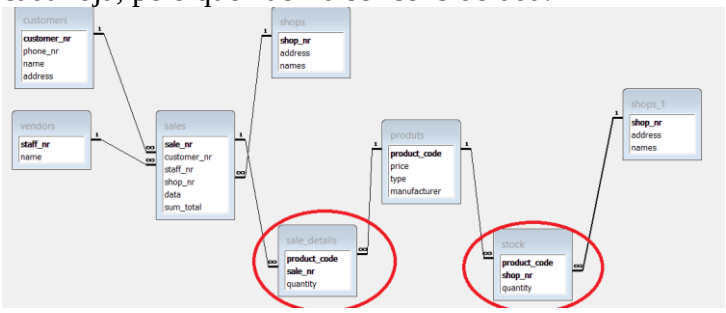
2.b) De seguida remova os caminhos múltiplos que eventualmente existam no esquema de base de dados. Na representação gráfica das ligações de 1:N, a tabela com uma única linha é desenhada em cima e a tabela com várias linhas é desenhada por baixo.

Resp: vamos utilizar uma tabela aliás "shop_1" obtendo assim uma estrutura arborescente

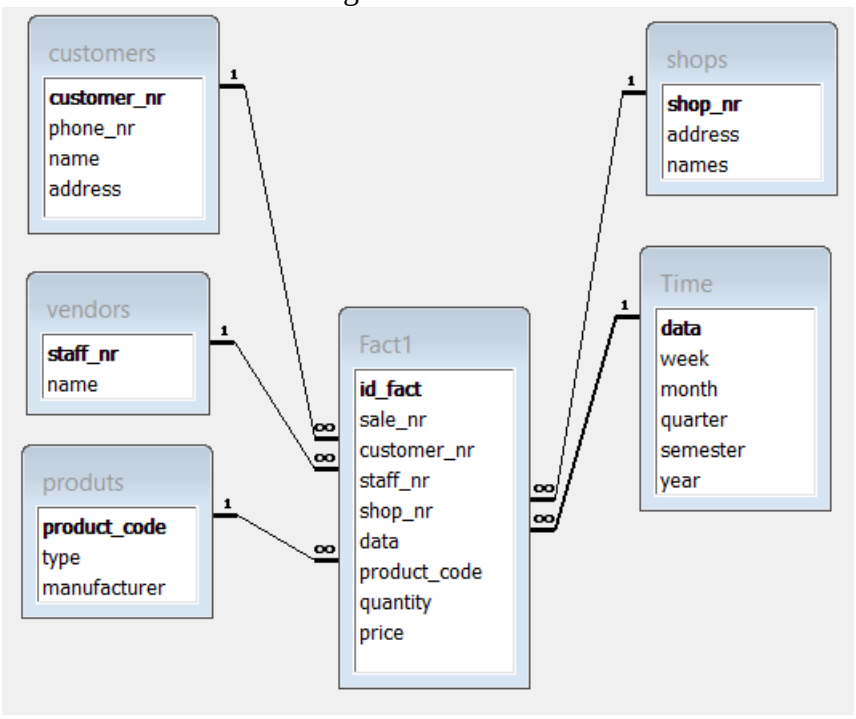


2.c) Pretendemos desenhar um "Data Warehouse" relacional em estrela ou em constelação, i.e. com duas ou mais estrelas com a granularidade dos detalhes das vendas. Defina a(s) tabela(s) de factos e mostre a tabela depois da desnormalização dos dados. Defina as dimensões com os níveis de agregação para 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.

Resp:
Visto que existem duas tabelas na parte de baixo da árvore, e pretendermos manter a granularidade dos detalhes das vendas, existem à partida 2 tabelas de factos. Contudo, a tabela Stock não regista os movimentos, mas a quantidade existente de cada produto em cada loja, pelo que não irá ser considerada.



O DW em estrela será o seguinte:



Critérios de correção 2.a, 2.b e 2.c (6 décimas):
- deve ser apresentado o esquema sem caminhos múltiplos e o DW com a tabela de factos
- a dimensão Tempo deve ser sempre criada
- para casa alínea, penalização de 1 a 2 décimas para erros ou omissões

2.d) Quais as tabelas da base de dados que foram desnormalizadas? Quais as tabelas da base de dados que não foram utilizadas no “Data Warehouse”?
Resp:
Foram desnormalizadas as tabelas Sales e Sales_details na tabela Fact1.

2.e) Crie duas perguntas e traduza para SQL utilizando pelo menos duas dimensões (OLAP).
Resp:
Consulta 1: Quantidades vendidas por cliente e por ano

```
TRANSFORM sum(Fact1.quantity) AS SomaDequantity
SELECT Fact1.customer_nr, sum(Fact1.quantity) AS [Total de quantity]
FROM Fact1
GROUP BY Fact1.customer_nr
PIVOT Format([data],"yyyy"
```

Fact1_Tabela de Referência Cruzada : Consulta Tabela de referência cruzada

customer_nr	Total de quantity	2014	2015
1	10232	4302	5930
2	10438		10438

Consulta: Quantidades vendidas por loja e vendedor

```
TRANSFORM sum(Fact1.quantity) AS SomaDequantity
SELECT Fact1.shop_nr, sum(Fact1.quantity) AS [Total de quantity]
FROM Fact1
GROUP BY Fact1.shop_nr
PIVOT Fact1.staff_nr;
```

Fact1_Tabela de Referência Cruzada1 : Consulta Tabela de referência cruzada

shop_nr	Total de quantity	staff_nr=1	staff_nr=2
1	10232	10232	
2	10438		10438

Critérios de correção 2.d e 2.e (4 décimas):
- devem ser referidas as tabelas que foram desnormalizadas
- devem ser apresentados os resultados e as consultas SQL com a utilização do Pivot

3) (1 valor) *Information Retrieval*
Escreva um texto, com pelo menos 500 palavras, onde descreva:

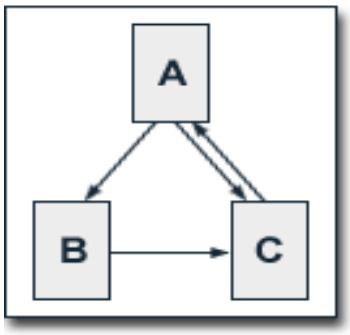
3.a) o que entende por PageRank

Resp:
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

Dada a dimensão da Web, existem métodos iterativos que permitem calcular o *PageRank* sem recorrer aos sistemas de equações.

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

O *PageRank* é o algoritmo que permite calcular o "valor" de uma página na Web. O valor da página não depende apenas da quantidade de *links* apontados para ela, mas do "valor" das páginas que apontam para ela. No exemplo PR(C) depende de PR(A) e PR(B), PR(C) = 0.5 + 0.5 (PR(A) / 2 + PR(B) /1).

3.b) o que entende por SEO (Search Engine Optimization)

Resp:
Um motor de pesquisa organizam a sua informação em grandes matrizes com palavras e documentos (ou páginas).

Palavras\Documentos	1	2	3	4	5	6	7	8	9	10	11	12	13	14
universidade	x	x	x	x			x				x	x		
aberta	x		x	x			x							
educação	x				x	x		x		x		x		x
tecnologias	x					x				x				x
curso		x	x					x	x		x			x
informática		x	x					x	x		x			x
gestão					x			x	x			x		x
cultura					x		x		x			x	x	x
matemática				x				x	x	x		x	x	x

Os motores de pesquisa têm três subsistemas:
i) Subsistemas de indexação (Documento/Página) que insere novas colunas na matriz com base nas palavras do documento;
ii) Subsistema Matriz (Documento, Palavra);
iii) Subsistemas de Pesquisa (Palavra) que para cada palavra, ou linha, verifica os documentos que estão associados;

A pesquisa por palavra utiliza a álgebra booleana, selecionando os documentos mais relevantes.

O Marketing na Web fez crescer uma nova disciplina o SEO, que tem como objetivo melhorar o desempenho de uma dada página nos motores de pesquisa, para uma ou mais palavras.

Tal como referimos nos motores de pesquisa existem palavras e páginas, existem também duas formas de otimização em SEO:
1) otimização "on-page" está relacionada com a escolha das palavras ou "keywords"
2) otimização "off-page" está relacionada com o *PageRank*

Na otimização "on-page" distinguem-se ainda dois métodos:
1.1) métodos "white hat" ou métodos com ética
1.2) métodos " black hat" ou métodos sem ética

Dos métodos "black hat" podemos referir:
- repetir a utilização de uma palavra para aumentar a sua relevância na página;
- utilizar texto invisível ao utilizador mas captadas pelo motor de busca (utilizar palavras escondidas em letras da mesma cor do fundo ou formatar o tamanho da letra para zero);
- uso de redireccionamentos não autorizados ou de camuflagem do verdadeiro conteúdo da página;

3.c) e a sua importância do PageRank no SEO

A importância do PageRank no SEO é evidente na otimização “off-page” do website.

A otimização “off-page” passa por criar links externos (“link building”) que façam referência ao website, como por exemplo:
- escrever num blog e/ou fóruns sobre o website
- referir o website nas redes sociais
- submeter o website em motores de pesquisa (Google)
- submeter o website em diretórios (Yahoo)

Critérios de Correção (10 décimas):
- explicação e exemplifique o algoritmo PageRank (4 décimas)
- explicação do SEO (3 décimas)
- a importância do SEO do PageRank (3 décimas)
- penalização de 1 a 2 décimas para erros ou omissões