e-Fólio B

U.C. 21103

Sistemas de Gestão de Bases de Dados

2019-2020

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. 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 objectividade 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

1) (1 valor) Capítulo 15, Concorrency Control
1.a) Defina o protocolo 2-PL. Quais as vantagens e desvantagens? Justifique a resposta.

1.b) Considere o protocolo 2-PL e explique detalhadamente a execução das seguintes transações, usando os operadores X-lock(_), S-lock(_) e Unlock(_). Como classifica a concorrência destas duas transações? Justifique a resposta.

	T1	T2
1	Read A	
2		Read B
3	Write A	
4		Read A
5		Write A
6		Write B
7	Read B	
8	Write B	

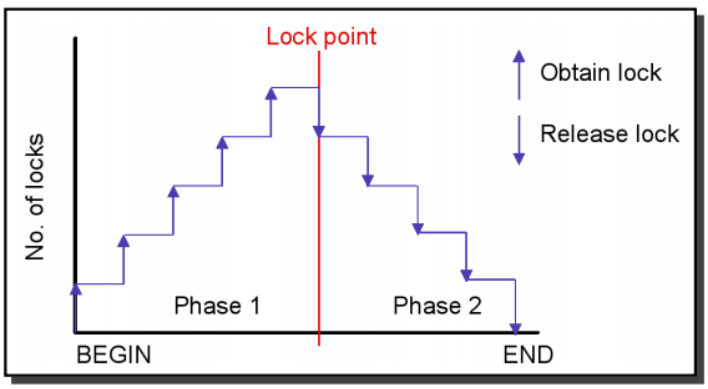
Resposta:

1. a) Defina o protocolo 2-PL. Quais as vantagens e desvantagens? Justifique a resposta.

O protocolo 2PL (2-phase locking) é um algoritmo de bloqueio utilizado para o controlo de concorrência entre transações. Os algoritmos de bloqueio são os mais utilizados nos SGBD e, entre eles, o 2PL é o mais aplicado.

O protocolo 2PL utiliza dois tipos de bloqueios (locks):
• Bloqueio partilhado (S-lock) (utilizado nas operações de leitura (Read)): o item de dados pode ser partilhado por várias transações.
• Bloqueio exclusivo (X-lock) (utilizado nas operações de escrita (Write)): o item de dados não pode ser partilhado por várias transações.

O protocolo 2PL exige que todas as transações solicitem todos os bloqueios que necessitem antes de libertar qualquer um dos bloqueios que detenha. Desta forma a gestão de bloqueios é realizada em duas fases:
• Fase de crescimento ou expansão: a transação apenas pode adquirir bloqueios.
• Fase de encolhimento ou contenção: a transação apenas pode libertar bloqueios.
O ponto de mudança de fases é designado de ponto de bloqueio ('lock point').



• Vantagens:
- produz escalonamentos serializáveis
- de fácil implementação

• Desvantagens:
- os pedidos de ‘locks’ são problemáticos para o ‘lock manager’
- limita a concorrência, colocando em transações em fila de espera
- não evita a ocorrência de ‘deadlock’
- não evita recuperações em cascata (‘cascading rollback’)

1.b) Considere o protocolo 2-PL e explique detalhadamente a execução das seguintes transações, usando os operadores X-lock(_), S-lock(_) e Unlock(_). Como classifica a concorrência destas duas transações? Justifique a resposta.

T1	**T2**	**Comentários**
S-lock(A), R(A)		
	S-lock(B), R(B)	
X-lock(A), W(A)		
	Enqueue (A)	Pretende-se R(A) Não é possível T2 fazer X-lock(A) Deve aguardar
Enqueue(B)		Pretende-se R(B) e W(B) Não é possível T1 fazer X-lock(B) Deve aguardar
		T1 e T2 aguardam a libertação de recursos, uma da outra. Estamos na presença de 'Deadlock'

O sequenciamento não é serializável e não pode ocorrer utilizando o protocolo 2PL, o 'deadlock' ocorre.

Critérios de correção:
- 1.a) 2 décimas, definir 2-PL, vantagens e desvantagens (igual ao p-fólio anterior)
- 1.b) 4 décimas, tabela com transações e locks
- 1.b) 4 décimas, identificar 'deadlock'
- erros, omissões ou redundâncias: -20% a -100%

2) (1 valor) Capítulo 16, Recovery System

2.a) Quais as principais fases que devem ser consideradas na recuperação? Justifique a resposta.

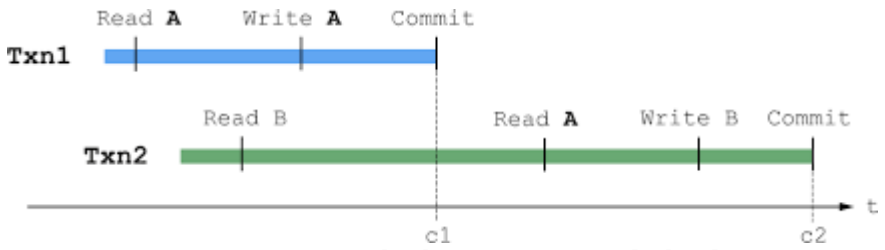


Figura: exemplo de transações na linha do tempo

2.b) Considere a seguinte sequência de log de duas transações em uma conta bancária, com saldo inicial de 12.000, que transfere 2.000 para um pagamento e recebe juros de 5%. Aplique o algoritmo de recuperação ao seguinte log. Represente as transações na linha do tempo como na figura em cima e acrescente os registos na recuperação. Justifique a resposta.

300. Checkpoint
310. T1 start
320. T1 B old=12000 new=10000
330. T1 M old=0 new=2000
340. T1 commit
350. T2 start
360. T2 B old=10000 new=10500
FAIL

Resposta:

2.a) Quais as principais fases que devem ser consideradas na recuperação? Justifique a resposta.

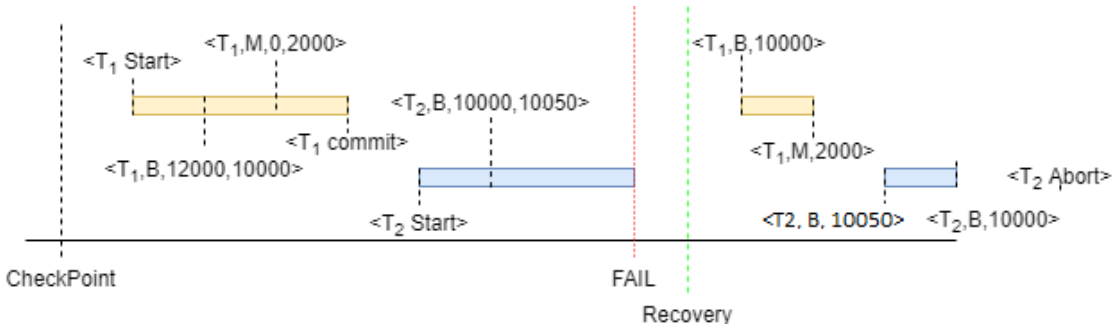
As fases do protocolo de recuperação são: refazer (redo phase) e desfazer (undo phase).

A fase de refazer (Redo) percorre o *log* desde o último *checkpoint*, refazendo modificações.
1- A lista de transações a serem recuperadas, Undo-list, é inicializada com a lista de transações do checkpoint.
2- Sempre que um registro de log normal na forma <Ti, Xj, V1, V2> ou apenas um registro de log Redo da forma <Ti, Xj, V2>, a operação é refeita.
3- Sempre que um registro de log do tipo <Ti start> for encontrado, Ti é adicionado ao Undo-list.
4- Sempre que encontramos um registo na forma <*Ti* abort> ou <*Ti* commit>, *Ti* é removido da Undo-list.

A fase de desfazer (undo phase) percorre o *log* da entrada mais recente para a mais antiga.
1- Sempre que encontrar um registro de uma transação na Undo-list, ele executa o Undo como se o registro tivesse sido encontrado durante a reversão de uma transação falhada.
2- Quando encontra um registo na forma <Ti start> remove a transação Ti na undo-list, ele escreve <Ti abort> e remove a transação da undo-list.
3- A fase Undo termina assim que a Undo-list esteja vazia.

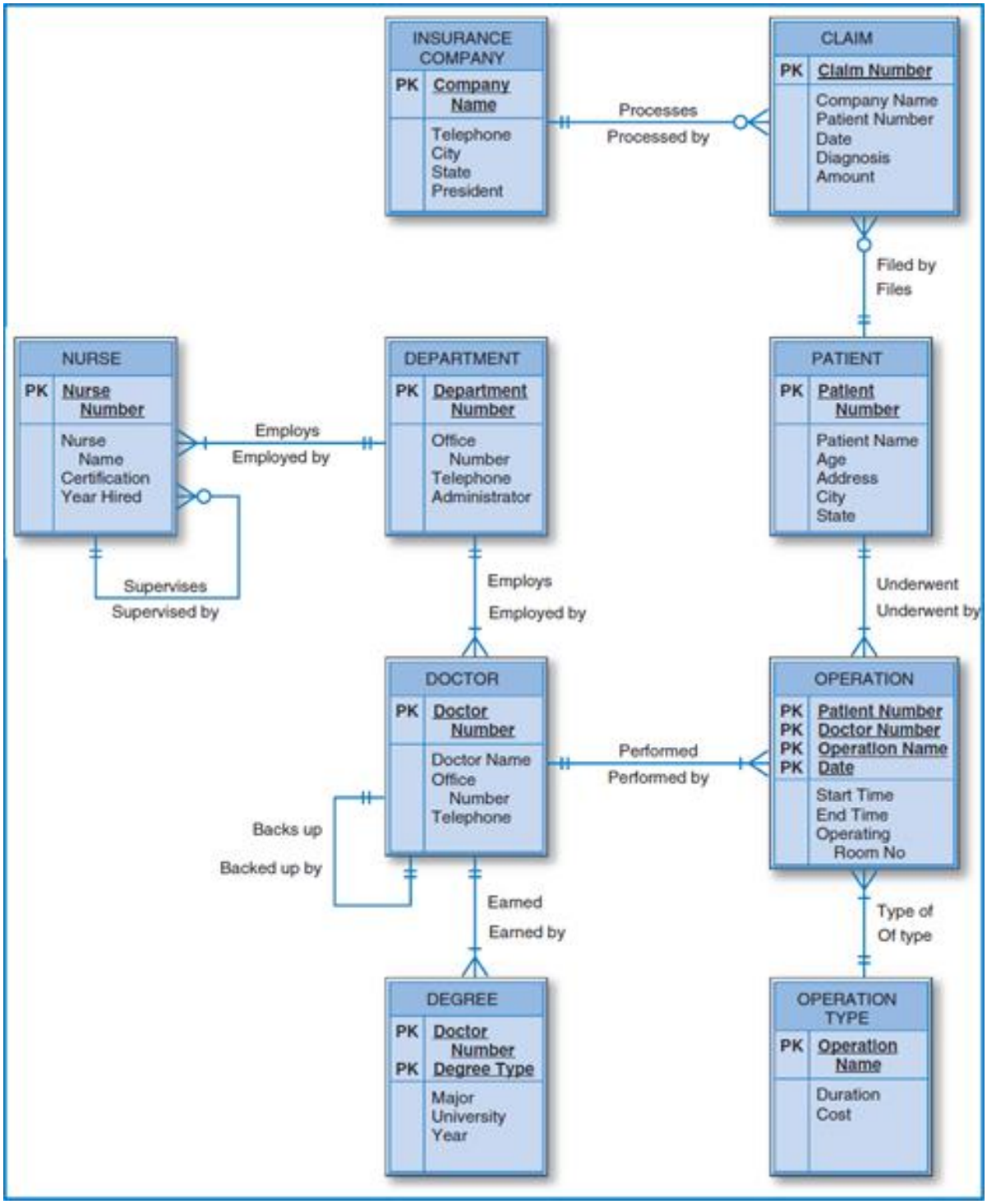
2.b) Represente as transações na linha do tempo como na figura em cima e acrescente os registos na recuperação. Justifique a resposta.

300. Checkpoint 310. T1 start 320. T1 B old=12000 new=10000 330. T1 M old=0 new=2000 340. T1 commit 350. T2 start 360. T2 B old=10000 new=10500 FAIL ……. 370. T1, B, 10000 (redo) 380. T1, M, 2000 (redo) 390. T2, B, 10050 (redo) 400. T2, B, 10000 (undo) 410. T2 abort (undo)	Início Redo-Phase Undo-list= T1 Redo T1, B, 10000 Redo T1, M, 2000 Undo-list =vazia Undo-list =T2 Redo T2, B,10500 Fim Redo-Phase …….	Fim Undo-Phase Undo-list= vazia Undo T2, B, 10000 Início Undo-Phase ……. Registos acrescentados na Redo-Phase e Undo-Phase



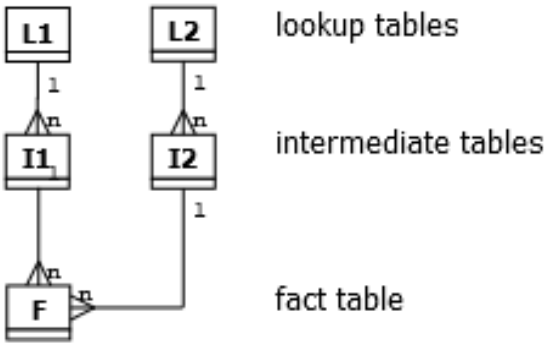
Critérios de correção:
- 2.a) 2 décimas as 2 fases de recuperação (igual ao p-fólio anterior)
- 2.b) 3 décimas, fase redo
- 2.b) 3 décimas, fase undo
- 2.b) 2 décimas, representação no tempo
- erros, omissões ou redundâncias: -20% a -100%

Para as perguntas **3) e 4) Desnormalização e Data Warehousing**, considere a seguinte base de dados:



3) (1 valor) Desnormalização

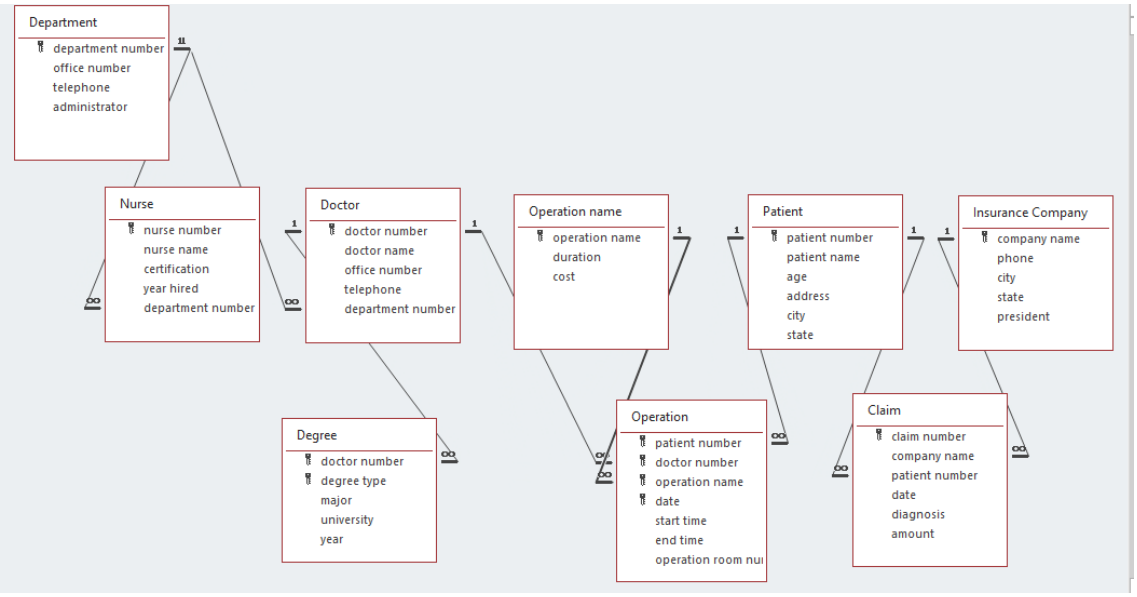
3.a) Reutilize a base de dados transacional na 3ª forma normal. Faça o carregamento de dados. Represente graficamente as ligações de 1:N, a tabela com uma única linha é desenhada em cima e a tabela com várias linhas é desenhada por baixo. Depois de representar as tabelas classifique-as segundo a tipologia indicada.



3.b). Encontre a 1FD (1ª forma desnormalizada) e a 2FD (1ª forma desnormalizada). Justifique a resposta.

Resposta:

3.a) Base de dados com a representação gráfica pedida.



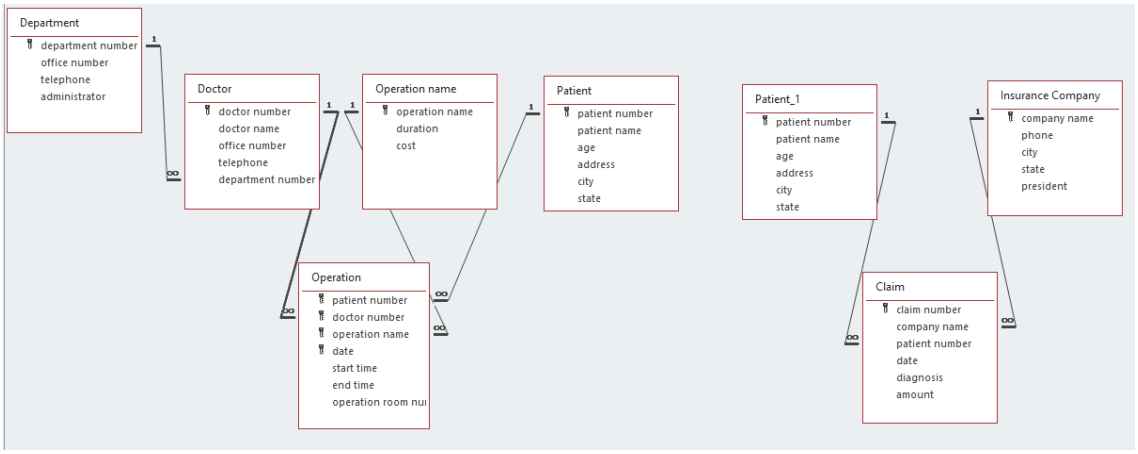
A tipologia das tabelas é a seguinte:

tabela	**tipologia**
claim	facts
degree	facts
nurse	facts
operation	facts
doctor	intermediate
departament	lookup
insurance company	lookup
operation name	lookup
patient	lookup

3.b) Encontre a 1FD (1ª forma desnormalizada) e a 2FD (1ª forma desnormalizada). Justifique a resposta.

1FD - a base de dados está na 1FD visto que não tem caminhos múltiplos

2FD - Vamos desprezar as tabelas de factos Nurse e Degree. Para as tabelas de factos Operation e Claim temos a seguinte 2FD.



Critérios de correção:
- 3.a) 5 décimas, base dados na 3FN e tipologia das tabelas
- 3.b) 5 décimas, 1 e 2 DF
- erros, omissões, redundâncias ou indentação desadequada: -20% a -100%

4) (1 valor) Data Warehouse

4.a) Pretendemos desenhar um "Data Warehouse" relacional em estrela ou em constelação, i.e. com duas ou mais estrelas com a maior granularidade possível. 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. Quantas tabelas de factos encontrou? Preencha a 'bus matrix' (ou 'business matrix') apresentada em baixo. Justifique a resposta.

		dimension	dimension	dimension	dimension
business process	fact table	1	2	3	4
Process X	A	X			X
	B		X	X	
	C				X

4.b) Crie duas perguntas e traduza para SQL com Pivot Tables utilizando pelo menos duas dimensões (OLAP). Justifique a resposta.

Resposta:

4.a) 'Bus matrix'

		doctor	operation name	patient	insurance company	date
database	fact table	dimension	dimension	dimension	dimension	dimension
hospital	1. operation	X	X	X		X
	2. claim			X	X	X

4.b) Perguntas OLAP (resposta parcial)
4.b.1) Quantas e que operações foram realizadas por que médicos?

```
TRANSFORM Count(Operation.patient_number)
SELECT Operation.operation_name, Count(Operation.patient_number)
FROM Operation
GROUP BY Operation.operation_name
PIVOT Operation.doctor_number
```

	doctors		
operation name	**a**	**b**	**c**
x	16	4	
y		31	11
z			51

Critérios de correção:
- 4.a) 6 décimas, 'bus matrix' com 5 dimensões
- 4.b) 4 décimas, 2 perguntas OLAP com referências cruzadas (A versus B)
- erros, omissões, redundâncias ou indentação desadequada: -20% a -100%

5. (1 valor) Information Retrieval

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:

$$TF(d, t) = \log_e\left(1 + \frac{n(d, t)}{n(d)}\right)$$

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:

$$IDF(t) = \log_{10}\left(\frac{N}{n(t)}\right)$$

Assim, a relevância de um termo "t" num documento "d" é dado por:

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

Para a seguinte tabela de frequências de termos versus documentos, encontre para cada documento o termo mais relevante.

Termos \ Documentos	1	2	3	4	5	6	7	8	9	10	11	12	13	14
universidade	81	70	0	72	0	0	60	224	200	0	240	112	96	40
aberta	0	40	100	135	48	0	16	54	90	90	0	6	30	0
informática	90	24	64	224	0	180	0	72	48	0	120	4	70	243
gestão	0	36	0	98	300	630	48	36	72	0	40	0	90	240
humanidades	360	189	80	216	0	120	160	150	98	20	180	72	36	120
matemática	360	10	175	0	540	120	560	9	160	9	420	160	80	0
ambiente	350	147	16	144	0	0	0	0	504	240	0	0	0	324
educação	0	216	0	60	105	648	0	96	240	0	40	49	2	0
história	56	0	250	0	0	720	120	72	81	96	147	24	180	20

Para um determinado termo "t" será possível encontrar os documentos mais relevantes? Discuta a abordagem TF-IDF. Justifique a resposta.

Resposta:

Da tabela dada, (i) somamos as colunas e (ii) contamos as linhas com valores superiores a zero.

Palavras \ Documentos	1	2	3	4	5	6	7	8	9	10	11	12	13	14	count()	IDF
universidade	81	70	0	72	0	0	60	224	200	0	240	112	96	40	10	0,146
aberta	0	40	100	135	48	0	16	54	90	90	0	6	30	0	10	0,146
informática	90	24	64	224	0	180	0	72	48	0	120	4	70	243	11	0,105
gestão	0	36	0	98	300	630	48	36	72	0	40	0	90	240	10	0,146
humanidades	360	189	80	216	0	120	160	150	98	20	180	72	36	120	13	0,032
matemática	360	10	175	0	540	120	560	9	160	9	420	160	80	0	12	0,067
ambiente	350	147	16	144	0	0	0	0	504	240	0	0	0	324	7	0,301
educação	0	216	0	60	105	648	0	96	240	0	40	49	2	0	9	0,192
história	56	0	250	0	0	720	120	72	81	96	147	24	180	20	11	0,105
sum()	1297	732	685	949	993	2418	964	713	1493	455	1187	427	584	987		

TF.IDF	1	2	3	4	5	6	7	8	9	10	11	12	13	14
universidade	0,01	0,01	0,00	0,01	0,00	0,00	0,01	0,04	0,02	0,00	0,03	0,03	0,02	0,01
aberta	0,00	0,01	0,02	0,02	0,01	0,00	0,00	0,01	0,01	0,03	0,00	0,00	0,01	0,00
informática	0,01	0,00	0,01	0,02	0,00	0,01	0,00	0,01	0,00	0,00	0,01	0,00	0,01	0,02
gestão	0,00	0,01	0,00	0,01	0,04	0,03	0,01	0,01	0,01	0,00	0,00	0,00	0,02	0,03
humanidades	0,01	0,01	0,00	0,01	0,00	0,00	0,00	0,01	0,00	0,00	0,00	0,01	0,00	0,00
matemática	0,02	0,00	0,02	0,00	0,03	0,00	0,03	0,00	0,01	0,00	0,02	0,02	0,01	0,00
ambiente	0,07	0,06	0,01	0,04	0,00	0,00	0,00	0,00	0,09	0,13	0,00	0,00	0,00	0,09
educação	0,00	0,05	0,00	0,01	0,02	0,05	0,00	0,02	0,03	0,00	0,01	0,02	0,00	0,00
história	0,00	0,00	0,03	0,00	0,00	0,03	0,01	0,01	0,01	0,02	0,01	0,01	0,03	0,00

A frequência do termo 'ambiente' é dada por:
TF('ambiente', d10) = ln(1 + 240 / 455)

A frequência inversa do documento é uma medida de quanta informação a palavra fornece, isto é, se é comum ou rara em todos os documentos. Quanto mais rara é a palavra, maior é o IDF (inverse document frequency):
IDF ('ambiente') = $\log_{10}(14 / 7) = 0,301$

O TF.IDF é dado pelo produto das duas métricas, realçando uma alta frequência de termo no documento fornecido e uma baixa frequência de documento do termo em toda a coleção de documentos:
TF.IDF('ambiente', d1) = 240/455 * 0,301= 0,13

O termo 'ambiente' apresenta um maior TD.IDF no conjunto de documentos.