Cento Universitário UNA

Sistemas de Informação

Recuperação de Informação

Práticas de Laboratório

Wesley Dias Maciel

2019/01

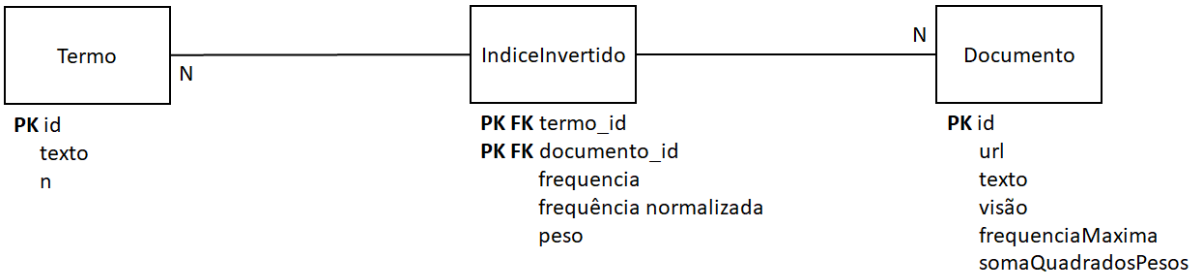
Centro Universitário UNA
Sistemas de Informação
Recuperação de Informação
Prática de Laboratório
Wesley Dias Maciel
2019/01

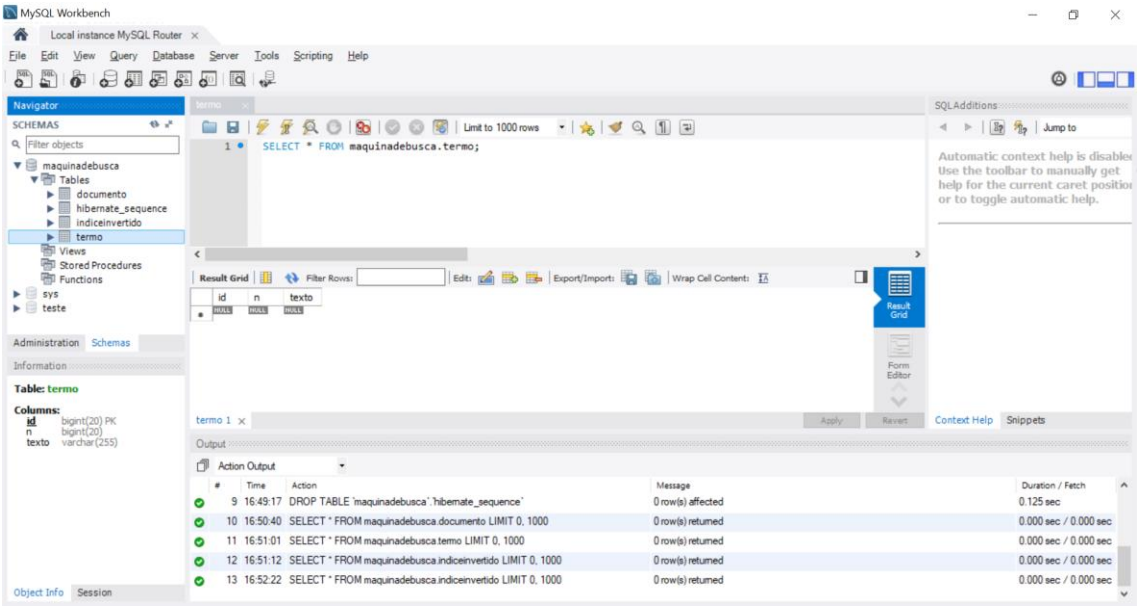
Indexador

Centro Universitário UNA
Sistemas de Informação
Recuperação de Informação
Prática de Laboratório
Wesley Dias Maciel
2019/01

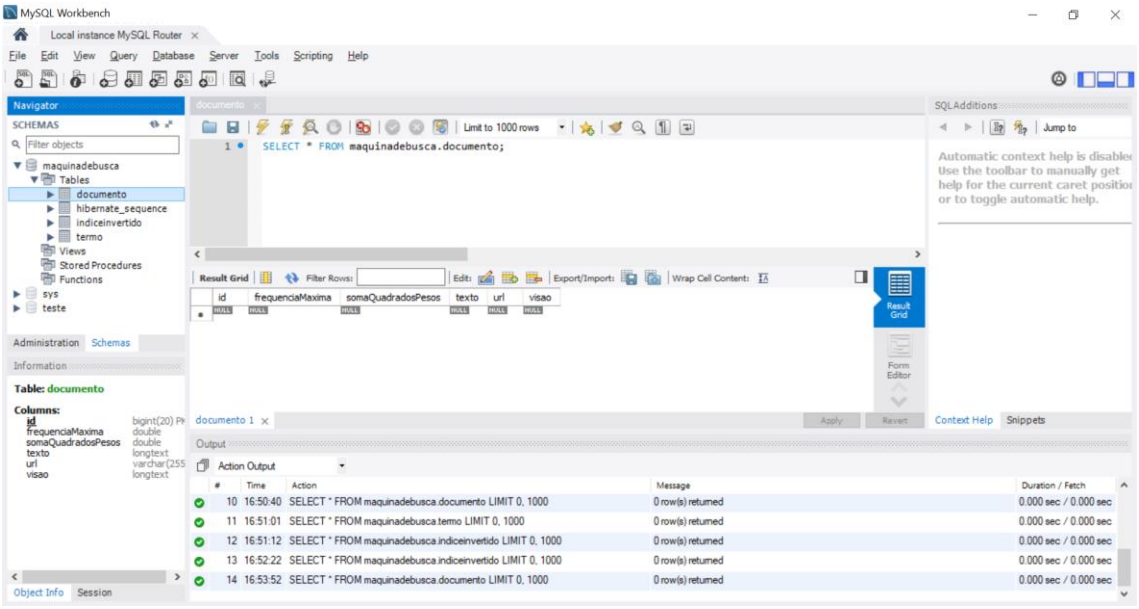
Prática 16

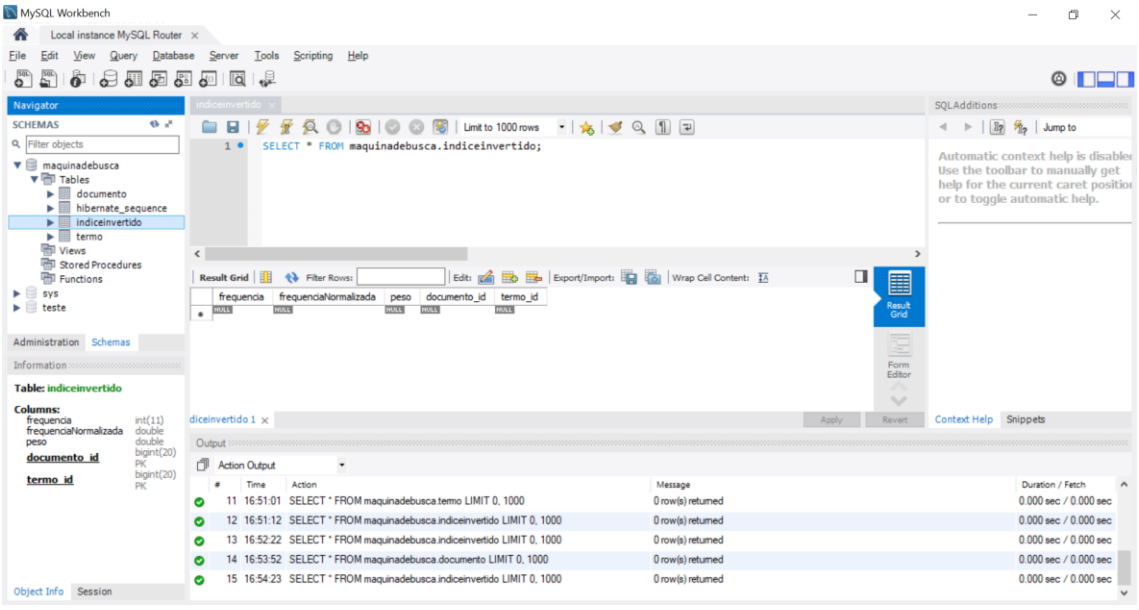
1) Você está recebendo, juntamente com esta prática, uma planilha Excel com um exemplo de cálculo de pesos no modelo de espaço vetorial e também o projeto da aplicação. Nesta versão, o projeto inicia a implementação do indexador. O projeto cria as tabelas Termo, IndiceInvertido e Documento no banco de dados como apresentado nas figuras abaixo:





Centro Universitário UNA
Sistemas de Informação
Recuperação de Informação
Prática de Laboratório
Wesley Dias Maciel
2019/01





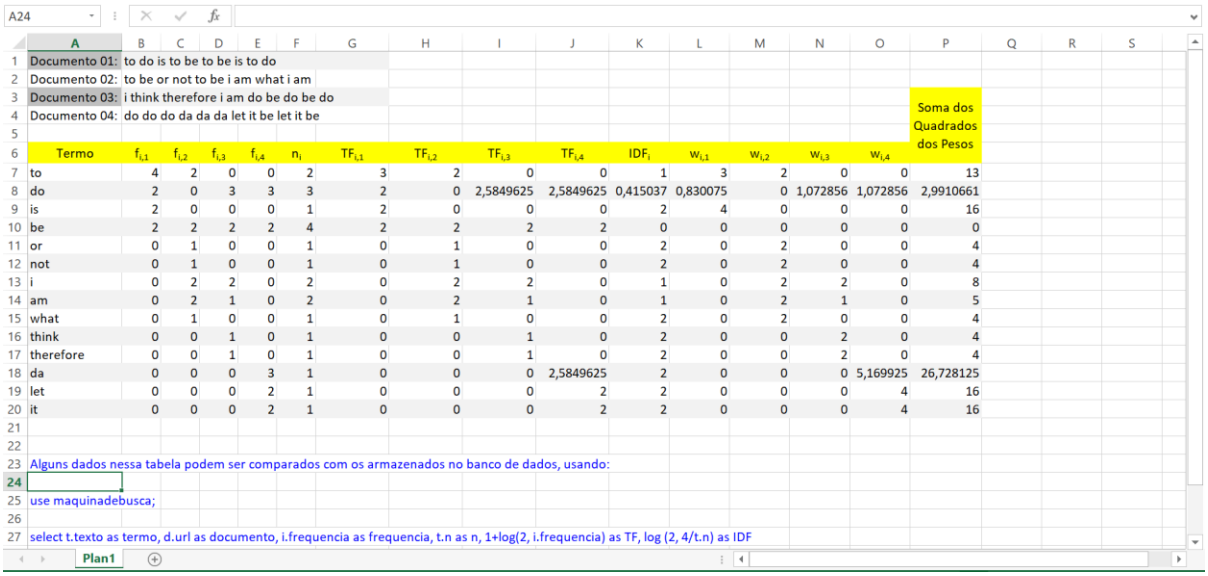
Nesta versão, os documentos usados são os apresentados nos slides empregados na apresentação do conteúdo. Analise o código do projeto. Execute o projeto usando o Postman. Em seguida, execute o comando abaixo no MySQL e verifique se as frequências estão sendo calculadas corretamente.

Centro Universitário UNA
Sistemas de Informação
Recuperação de Informação
Prática de Laboratório
Wesley Dias Maciel
2019/01

```
use maquinadebusca;

select t.texto as termo, d.url as documento, i.frequencia as frequencia
from termo t, documento d, indiceinvertido i
where t.id = i.termo_id and i.documento_id = d.id
order by t.texto, d.url;
```

Planilha Excel disponibilizada juntamente com a prática:



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Documento 01:	to do is to be to be is to do														
2	Documento 02:	to be or not to be i am what i am														
3	Documento 03:	i think therefore i am do be do be do														Soma dos Quadrados dos Pesos
4	Documento 04:	do do do da da da let it be let it be														
5																
6	Termo	$f_{i,1}$	$f_{i,2}$	$f_{i,3}$	$f_{i,4}$	n_i	$TF_{i,1}$	$TF_{i,2}$	$TF_{i,3}$	$TF_{i,4}$	IDF_i	$w_{i,1}$	$w_{i,2}$	$w_{i,3}$	$w_{i,4}$	
7	to	4	2	0	0	2	3	2	0	0	1	3	2	0	0	13
8	do	2	0	3	3	3	2	0	2,5849625	2,5849625	0,415037	0,830075	0	1,072856	1,072856	2,9910661
9	is	2	0	0	0	1	2	0	0	0	2	4	0	0	0	16
10	be	2	2	2	2	4	2	2	2	2	0	0	0	0	0	0
11	or	0	1	0	0	1	0	1	0	0	2	0	2	0	0	4
12	not	0	1	0	0	1	0	1	0	0	2	0	2	0	0	4
13	i	0	2	2	0	2	0	2	2	0	1	0	2	2	0	8
14	am	0	2	1	0	2	0	2	1	0	1	0	2	1	0	5
15	what	0	1	0	0	1	0	1	0	0	2	0	2	0	0	4
16	think	0	0	1	0	1	0	0	1	0	2	0	0	2	0	4
17	therefore	0	0	1	0	1	0	0	1	0	2	0	0	2	0	4
18	da	0	0	0	3	1	0	0	0	2,5849625	2	0	0	0	5,169925	26,728125
19	let	0	0	0	2	1	0	0	0	0	2	2	0	0	0	4	16
20	it	0	0	0	2	1	0	0	0	0	2	2	0	0	0	4	16
21																
22																
23	Alguns dados nessa tabela podem ser comparados com os armazenados no banco de dados, usando:															
24																
25	use maquinadebusca;															
26																
27	select t.texto as termo, d.url as documento, i.frequencia as frequencia, t.n as n, 1+log(2, i.frequencia) as TF, log (2, 4/t.n) as IDF															

Plan1

2) Altere o projeto, para que ele calcule a quantidade de documentos em que cada termo i ocorre (n_i) e armazene essa quantidade na coluna correspondente da tabela Termo.

3) Altere o projeto, para que ele calcule a frequência normalizada dos termos em cada documento e as armazene na coluna correspondente da tabela IndiceInvertido.

   OBS: frequenciaNormalizada = $frequencia_{i,j}$ / $frequenciaMaxima_j$

4) Altere o projeto, para que ele calcule os pesos dos termos em cada documento e os armazene na coluna correspondente da tabela IndiceInvertido.

   OBS:
   peso = $TF_{i,j} \times IDF_i$
   $TF_{i,j} = 1 + \log frequencia_{i,j}$
   $IDF_i = \log N / n_i$

5) Altere o projeto, para que ele calcule a soma dos quadrados dos pesos e os armazene na coluna correspondente da tabela Documento.

6) Realize testes com o seu projeto, usando os exercícios passados em sala de aula como base.

7) Altere o projeto, para que ele indexe os documentos coletados por seu coletor.

8) No projeto, sempre retornar respostas que obedeçam os códigos adequados do protocolo HTTP.

9) Analise a API do seu projeto. Sempre que necessário, faça alterações para melhoria da API, adequando-a ao padrão arquitetural REST (Representational State Transfer).

Lista de códigos de status HTTP:

1xx Informativa
- 100 Continuar
- 101 Mudando protocolos
- 102 Processamento (WebDAV) (RFC 2518)
- 122 Pedido-URI muito longo

2xx Sucesso
- 200 OK
- 201 Criado
- 202 Aceito
- 203 não-autorizado (desde HTTP/1.1)
- 204 Nenhum conteúdo
- 205 Reset
- 206 Conteúdo parcial
- 207-Status Multi (WebDAV) (RFC 4918)

3xx Redirecionamento
- 300 Múltipla escolha
- 301 Movido
- 302 Encontrado
- 303 Consulte Outros
- 304 Não modificado
- 305 Use Proxy (desde HTTP/1.1)
- 306 Proxy Switch
- 307 Redirecionamento temporário (desde HTTP/1.1)
- 308 Redirecionamento permanente (RFC 7538[2])

4xx Erro de cliente
- 400 Requisição inválida
- 401 Não autorizado
- 402 Pagamento necessário
- 403 Proibido
- 404 Não encontrado
- 405 Método não permitido
- 406 Não Aceitável
- 407 Autenticação de proxy necessária
- 408 Tempo de requisição esgotou (Timeout)

409 Conflito
410 Gone
411 comprimento necessário
412 Pré-condição falhou
413 Entidade de solicitação muito grande
414 Pedido-URI Too Long
415 Tipo de mídia não suportado
416 Solicitada de Faixa Não Satisfatória
417 Falha na expectativa
418 Eu sou um bule de chá
422 Entidade improcessável (WebDAV) (RFC 4918)
423 Fechado (WebDAV) (RFC 4918)
424 Falha de Dependência (WebDAV) (RFC 4918)
425 coleção não ordenada (RFC 3648)
426 Upgrade Obrigatório (RFC 2817)
450 bloqueados pelo Controle de Pais do Windows
499 cliente fechou Pedido (utilizado em ERPs/VPSA)

5xx outros erros (erro de servidor)

500 Erro interno do servidor (Internal Server Error)
501 Não implementado (Not implemented)
502 Bad Gateway
503 Serviço indisponível (Service Unavailable)
504 Gateway Time-Out
505 HTTP Version not supported