Calculando TF-IDF

implementação do Índice Remissivo

Suposições para esse exemplo

1. Considere um livro com três páginas;
2. Todo o texto será considerado;
3. Cada página terá os seguintes textos
    - Página 1: Maria vai para praia tomar sol.
    - Página 2: Na praia havia uma barraca e uma cadeira .
    - Página 3: Praia e Sol combina com domingo.

Os passos

1. Limpeza dos dados
2. Calcule a frequência das palavras, nas páginas e na obra;
3. Encontre o TF das palavras;
4. Encontre o IDF das palavras;
5. Encontre o TF-IDF das palavras;

Passo 1: Limpeza

Palavra
maria
vai
para
praia
tomar
sol
na

Passo 2: frequência

Palavra	Frequência
maria	1
vai	1
para	1
praia	3
tomar	1
sol	2
na	1
[illegible]	1

Passo 3: Encontrar o TF das palavras nas páginas

de ocorrências da palavra na página dividido pelo # de palavras na página

Página 1: Maria vai para praia tomar sol

Palavra	TF
maria	0.16
vai	0.16
para	0.16
praia	0.16
tomar	0.16

Passo 3: TF em todo o texto

Palavra	Frequência	Página 1	Página 2	Página 3
maria	1	0.16	0.0	0.0
vai	1	0.16	0.0	0.0
para	1	0.16	0.0	0.0
praia	3	0.16	0.12	0.16
tomar	1	0.16	0.0	0.0
sol	2	0.16	0.0	0.16
na	1	0.0	0.12	0.0
[illegible]	1	0.0	0.12	0.0

Passo 4: Encontrar IDF das palavras

IDF = $ln(#_de_páginas/#_de_páginas_que_contém_palavra) $

Palavra	IDF
maria	$ln(3/1) = 1.09$
vai	1.09
para	1.09
praia	$ln(3/3) = 0$
tomar	1.09
sol	$ln(3/2) = 0.41$

Passo 5: Encontrar TF-IDF das palavras em cada página

IDF = $ln(#_de_páginas/#_de_páginas_que_contém_palavra) $

Palavra	Página 1	Página 2	Página 3
maria	$0.16 \times 1.09 = 0.1744$		
vai	0.1744		
para	0.1744		
praia	$0.16 \times 0.0 = 0.0$	$0.12 \times 0.0 = 0.0$	$0.16 \times 0.0 = 0.0$
tomar	$0.16 \times 1.09 = 0.1744$		

Analise do TF-IDF Calculado

Palavra\pagina	maria	vai	para	praia	tomar	sol	na	havia
# 1	0.1744	0.1744	0.1744	0	0.1744	0.0656	0	0
# 2	0	0	0	0	0	0	0.1308	0.1308
# 3	0	0	0	0	0		0	0

Critério de seleção por agrupamento

Palavra	Pagina 1	Pagina 2	Pagina 3	Limite superior	Limite inferior (75%)
B				0.1208	0.0906
barraca	0	0.1208	0		
C				0.1744	0.1308
cadeira	0	0.1208	0		
com	0	0	0.1744		
combina	0	0	0.1744		
D	0	0		0.1744	0.1308

Critério de seleção global

Palavra	Pagina 1	Pagina 2	Pagina 3	Limite superior	Limite inferior (85%)
				0.1744	0.1482
B					
barraca	0	0.1208	0		
C					
cadeira	0	0.1208	0		
com	0	0	0.1744		
combina	0	0	0.1744		

Obrigado!

boa sorte