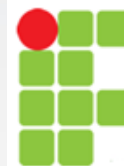




INSTITUTO FEDERAL
MINAS GERAIS

RECUPERAÇÃO DE INFORMAÇÃO

Profa. Mcs. Patrícia Proença
patricia.proenca@ifmg.edu.br



ATENÇÃO!!!

- ↓ O material a seguir é uma videoaula apresentada pela professora PATRÍCIA APARECIDA PROENÇA AVILA, como material pedagógico do IFMG, dentro de suas atividades curriculares ofertadas em ambiente virtual de aprendizagem. Seu uso, cópia e ou divulgação em parte ou no todo, por quaisquer meios existentes ou que vierem a ser desenvolvidos, somente poderá ser feito, mediante autorização expressa deste docente e do IFMG. Caso contrário, estarão sujeitos às penalidades legais vigentes”.
- ↓ Conforme Art. 2º§1º da Nota Técnica nº 1/2020/PROEN/Reitoria/IFMG (SEI 0605498, Processo nº 23208.002340/2020-04



INSTITUTO FEDERAL
MINAS GERAIS

Modelo Booleanocontinuando

Profa. Mcs. Patrícia Proença
patricia.proenca@ifmg.edu.br

Roteiro

- Ponderação IDF;
- Ponderação TF-IDF (*term-frequency – inverse document frequency*);
- Propriedades do TF-IDF;



Breve resumo da aula anterior!

Ponderação TF

- Uma variante da ponderação TF muito utilizada na literatura, pois torna os pesos diretamente comparáveis aos pesos IDF é a seguinte:

$$tf_{i,j} = \begin{cases} 1 + \log f_{i,j} & \text{se } f_{i,j} > 0 \\ 0 & \text{caso contrário} \end{cases}$$

- onde o logaritmo utiliza a base 2



Ponderação IDF

Ponderação IDF

- Alguns autores apresentaram trabalhos que mostraram:
 - ponderação pela frequência dos termos é útil para melhorar os resultados se comparada com a recuperação baseada em pesos binários;
 - mas o ganho ainda nem sempre é satisfatório;
 - Surgiu então estudos sobre melhorias na ponderação e surgiu a frequência inversa de documentos.

Ponderação IDF

- Sparck Jones desenvolveu uma **interpretação estatística da especificidade dos termos** (1972), chamada de **IDF**, que tornou-se a **pedra fundamental da ponderação de termos**;
- Essa interpretação tem uma base heurística que motivou várias pesquisas sobre abordagens que fornecessem um embasamento teórico para o IDF;
- Para entender a ponderação IDF, primeiramente é preciso estudar os conceitos de **exaustividade** e de **especificidade** dos termos da linguagem.

Ponderação IDF

- **Exaustividade:**
 - é uma propriedade das descrições dos **documentos**.
 - é interpretada como a abrangência que ela provê para os tópicos principais de um documento.
- Se adicionarmos novos termos do vocabulário a um documento, a exaustividade da descrição do documento aumenta.
- Além disso, a probabilidade que o documento satisfaça uma dada consulta também aumenta, isto é, a probabilidade de recuperação aumenta.

Ponderação IDF

- **Exaustividade:**
- Quanto **mais termos de indexação** são atribuídos a um documento, **mais exaustiva** fica sua descrição;
 - Sua probabilidade de recuperação em resposta a uma consulta selecionada aleatoriamente também aumenta.
- Problema: Se **muitos termos** forem atribuídos a um documento, ele irá ser retornado para **consultas** para as quais ele **não é relevante**.

Ponderação IDF

- Exaustividade:
- Isso sugere que o número médio de termos de indexação por documento deve ser otimizado de modo que a **probabilidade de relevância de um documento recuperado seja maximizada;**
- Esse número ótimo de termos de indexação define a **exaustividade ótima** para as descrições de tais documentos.

Ponderação IDF

- **Especificidade:**
 - é uma propriedade dos **termos de indexação**.
 - A especificidade de um termo de indexação é interpretada como quão bem um termo descreve o tópico de um documento.

Ponderação IDF

- **Especificidade:**

- é uma propriedade da semântica do termo, isto é, um termo é mais ou menos específico dependendo do seu significado.
- Exemplo: o termo “bebida” é menos específico do que os termos “chá” e “cerveja”.
- Se a indexação fosse feita manualmente, poderíamos esperar que o termo “bebida” fosse usado para indexar mais documentos do que os termos “chá” e “cerveja”.
- Uma alternativa é considerar a especificidade como uma função da utilização dos termos.

Ponderação IDF

- A **exaustividade** da descrição de um documento pode ser quantificada como o número de termos de indexação que ele possui;
- A **especificidade** de um termo é uma função do inverso do número de documentos nos quais ele ocorre.
 - Ou seja se um termo aparece muitas vezes quer dizer que ele não é muito específico de um documento.

Ponderação IDF

- Se as descrições dos documentos ficarem mais longas, a especificidade dos termos tende a ficar mais baixa.
 - Se um termo ocorrer em todos os documentos da coleção, sua especificidade é mínima e o termo não é útil para a recuperação.
- Ideia: ponderação de termos por especificidade/exaustividade
 - Para isso, os pesos dos termos podem ser representados como uma função das frequências relativas dos termos.

Ponderação IDF

- Como modelar os pesos usando os conceitos de especificidade e exaustividade?
- Problema: Com base nos conceitos de especificidade e exaustividade, gostaríamos de um modelo de ponderação que fizesse o seguinte:
 - 1. o valor do peso do termo será zero se ele puder ser encontrado em todos os documentos;
 - 2. o valor do peso do termo aumentará se ele estiver presente em poucos documentos.

Ponderação IDF

- Como modelar os pesos usando os conceitos de especificidade e exaustividade?
- 1) Verificar a ocorrência do termo k_i para cada documento d_j da coleção e armazenar em n_i
 - Quais documentos o termo aparece;
- 2) Calcular a frequência relativa inversa de cada termo N/n_i
 - Onde N é o total de documentos;
- 3) Aplicar a função log (na base 2) na frequência relativa inversa de cada termo.

$$IDF_i = \log \frac{N}{n_i}$$

- Assim, quando n_i se aproxima de N , temos que IDF_i se aproxima de zero.

Ponderação IDF - Exemplo

To do is to be.
To be is to do.

d_1

To be or not to be.
I am what I am.

d_2

I think therefore I am.
Do be do be do.

d_3

Do do do, da da da
Let it be, let it be.

d_4

Ponderação IDF - Exemplo

#	termo	n_i	$IDF_i = \log(N/n_i)$
1	to	2	1
2	do	3	0,415
3	is	1	2
4	be	4	0
5	or	1	2
6	not	1	2
7	I	2	1
8	am	2	1
9	what	1	2
10	think	1	2
11	therefore	1	2
12	da	1	2
13	let	1	2
14	it	1	2

Ponderação IDF - Comentários

- Observe que os termos mais seletivos na coleção ocorrem em apenas um documento;
- Os menos seletivos ocorrem em todos os documentos;
- Em coleções reais de grandes proporções, espera-se que os termos mais seletivos sejam substantivos e grupos de substantivos;

Ponderação IDF - Comentários

- Os termos menos seletivos são geralmente artigos, conjunções e preposições, que são frequentemente chamadas de *stopwords*.
- Atualmente, a ponderação IDF fornece a base para os esquemas de ponderação modernos e é usada por quase todos os sistemas modernos de RI.



Ponderação TF-IDF

Ponderação TF-IDF

- Proposto por Salton e Yang (1973);
- Esquema de ponderação de termos mais popular entre os modelos de RI;
 - Combinam os fatores IDF e as frequências dos termos.
- Seja $w_{i,j}$ o peso do termo associado ao par (k_j, d_j) . Então, definimos:

$$w_{i,j} = \begin{cases} (1 + \log f_{i,j}) \times \log \frac{N}{n_i} & \text{se } f_{i,j} > 0 \\ 0 & \text{caso contrário} \end{cases}$$

- que é conhecida por esquema de ponderação TF-IDF.

Ponderação TF-IDF - Exemplo

To do is to be.
To be is to do.

d_1

To be or not to be.
I am what I am.

d_2

I think therefore I am.
Do be do be do.

d_3

Do do do, da da da
Let it be, let it be.

d_4

Ponderação TF-IDF - Exemplo

TERMO	$f_{i,1}$	$f_{i,2}$	$f_{i,3}$	$f_{i,4}$		$w_{i,1}$			
to	4	2	-	-		3			
do	2	-	3	3					
is	2	-	-	-					
be	2	2	2	2					
or	-	1	-	-					
not	-	1	-	-					
I	-	2	2	-					
am	-	2	1	-					
what	-	1	-	-					
think	-	-	1	-					
therefore	-	-	1	-					
da	-	-	-	3					
let	-	-	-	2					
it	-	-	-	2					
Tamanho do documento	10	11	10	12					

$$w_{i,j} = \begin{cases} (1 + \log f_{i,j}) \times \log \frac{N}{n_i} & \text{se } f_{i,j} > 0 \\ 0 & \text{caso contrário} \end{cases}$$

Termo to no doc 1

- * $\log f_{i,j} = \log(4) = 2$ (base 2)
- * $\log N/n_i = \log(4/2) = 1$ (base 2)
- * $w_{i,j} = (1+2)*1 = 3$

Exercício para praticar

TERMO	$f_{i,1}$	$f_{i,2}$	$f_{i,3}$	$f_{i,4}$		$w_{i,1}$	$w_{i,2}$	$w_{i,3}$	$w_{i,4}$
to	4	2	-	-		3	2	0	0
do	2	-	3	3					
is	2	-	-	-					
be	2	2	2	2					
or	-	1	-	-					
not	-	1	-	-					
I	-	2	2	-					
am	-	2	1	-					
what	-	1	-	-					
think	-	-	1	-					
therefore	-	-	1	-					
da	-	-	-	3					
let	-	-	-	2					
it	-	-	-	2					
Tamanho do documento	10	11	10	12					

**CALCULE O TF-IDF
PARA OS DEMAIS
TERMOS DA COLEÇÃO.**

Ponderação TF-IDF

- Termos mais frequentes dentro de um documento e termos mais raros possuem um peso TF-IDF maior;
- Embora simples, os pesos TF-IDF são bastante eficazes, especialmente, para coleções genéricas:
 - Coleção de documentos sobre a qual não temos nenhuma informação.




Comentários

No decorrer da aula vimos...

- Dado um conjunto de termos de indexação para uma coleção de documentos, **nem todos os termos** são igualmente úteis para descrever o conteúdo dos documentos;
- Métodos usuais para ponderar termos envolvem o estudo da frequência dos termos presentes nos documentos.

No decorrer da aula vimos...

- Ponderação TF:
 - Baseada na frequência dos termos;
- Ponderação IDF:
 - Baseada na frequência relativa (inversa) dos termos;
- Ponderação TF-IDF:
 - Baseada em uma mescla entre a frequência dos termos e a frequência relativa (inversa) dos termos.

A wooden tray with various fruits and a glass of juice on a bed. The tray contains a bowl of mixed berries, a glass of juice, and some other fruits. The background is a light blue fabric, possibly a bedsheet. The text is overlaid on the image in white, bold, uppercase letters.

**QUE A GENTE SEMPRE
CARREGUE FÉ, AMOR
E PENSAMENTO
POSITIVO EM
NOSSOS DIAS.
BOA SEMANA!**