Universidade Federal do Rio de Janeiro (UFRJ)
Departamento de Ciência da Computação (DCC)

DEPARTAMENTO DE CIÊNCIA DA COMPUTAÇÃO - UFRJ -

Recuperação da Informação (MAB605)

Ponderação de Termos

Profa. Giseli Rabello Lopes

Roteiro

- Introdução
- Ponderação de termos
  - Ponderação da frequência de termos
  - Ponderação pela frequência inversa de documentos
  - Ponderação TF-IDF
  - Variantes do TF-IDF
  - Propriedades do TF-IDF
- Referências

Ponderação dos termos

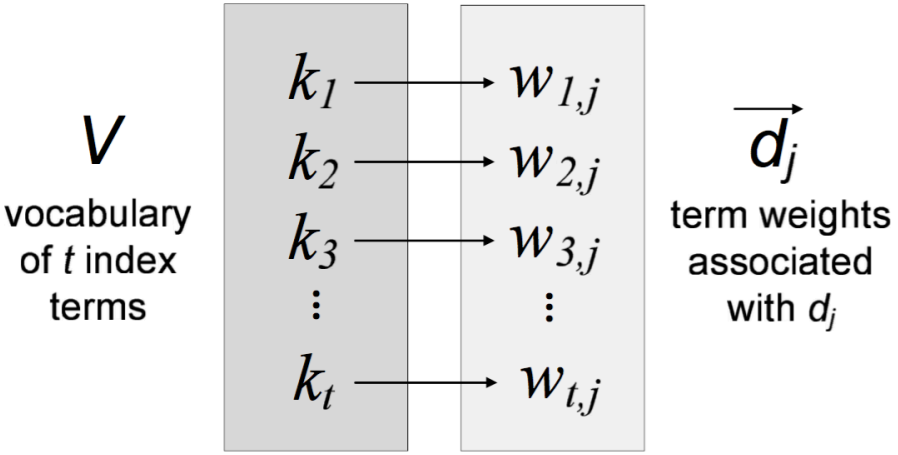
- Os termos de um documento não são igualmente úteis para descrever o conteúdo do documento
- Na verdade, existem termos de indexação que são simplesmente mais vagos do que outros
- Existem propriedades de um termo de indexação que são úteis para avaliar sua importância em um documento
  - Por ex., uma palavra que aparece em todos os documentos de uma coleção é completamente inútil para tarefas de recuperação

Ponderação dos termos

- Para caracterizar a importância de um termo, é associado um peso $w_{i,j} > 0$ para cada termo k_i que ocorre em um documento d_j
  - Se k_i não aparece em um documento d_j, então $w_{i,j} = 0$
- O peso $w_{i,j}$ quantifica a importância do termo de indexação k_i para descrever o conteúdo do documento d_j
- Esses pesos são úteis para computar um grau numérico (*rank*) para cada documento da coleção em relação a uma dada consulta

Ponderação dos termos

- Seja,
  - k_i um termo de indexação e d_j um documento
  - $V = \{k_1, k_2, \ldots, k_t\}$ o conjunto de todos os termos de indexação
  - $w_{i,j} \geq 0$ o peso associado com (k_i, d_j)
- Então, $\vec{d_j} = (w_{1,j}, w_{2,j}, \ldots, w_{t,j})$ como um vetor de pesos que contém o peso $w_{i,j}$ de cada termo $k_i \in V$ no documento d_j



Ponderação dos termos

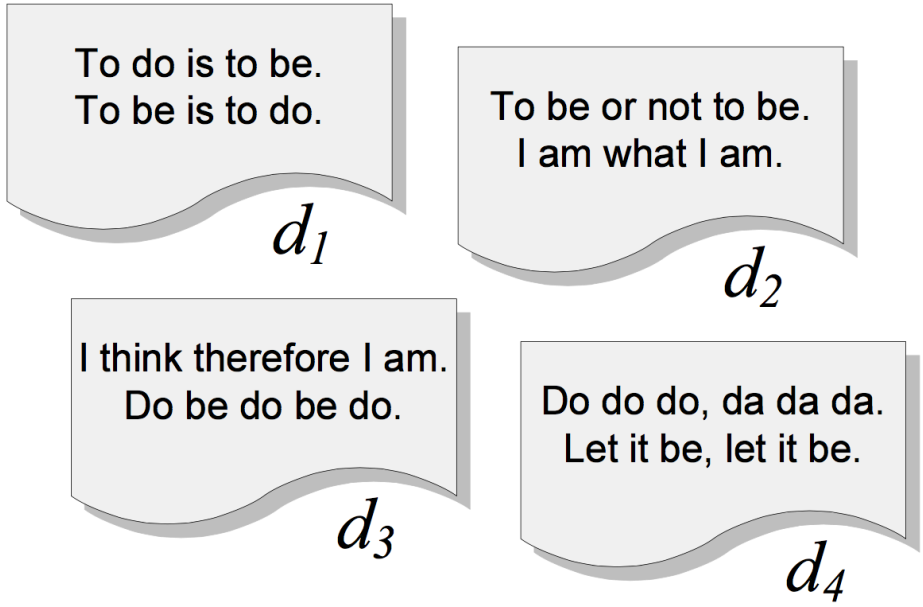
- Os pesos $w_{i,j}$ podem ser computados usando as **frequências de ocorrência** dos termos nos documentos
- Seja $f_{i,j}$ a frequência de ocorrência de um termo de indexação k_i no documento d_j
- A **frequência total de ocorrência** F_i do termo k_i na coleção é definida como

$$F_i = \sum_{j=1}^{N} f_{i,j}$$

  - onde N é o número de documentos da coleção

Ponderação dos termos

- A **frequência de documento** n_i (ou df_i) para um termo k_i é o número de documentos nos quais ele ocorre
  - Note que $n_i \leq F_i$
- Por ex., na coleção de documentos abaixo, os valores de $f_{i,j}$, F_i e n_i associados ao termo *do* são:
  - $f(do, d_1) = 2$
  - $f(do, d_2) = 0$
  - $f(do, d_3) = 3$
  - $f(do, d_4) = 3$
  - $F(do) = 8$
  - $n(do) = 3$



Ponderação TF-IDF

- Esquema de ponderação de termos TF-IDF:
  - Fundamentos do esquema de ponderação mais popular em RI
    - *Term frequency* (TF)
    - *Inverse document frequency* (IDF)

Ponderação da frequência dos termos

- **Hipótese de Luhn**. O valor (peso) $w_{i,j}$ é proporcional à frequência do termo $f_{i,j}$
  - Isto é, quanto mais frequentemente um termo k_i ocorrer no documento d_j maior será a sua frequência de termo $tf_{i,j}$
- Baseado na observação que termos com alta frequência são importantes para descrever os tópicos-chave de um documento
- Leva diretamente à seguinte formulação da ponderação TF:

$$tf_{i,j} = f_{i,j}$$

Ponderação da frequência dos termos

- Uma variante do TF utilizada na literatura é

$$tf_{i,j} = \begin{cases} 1 + \log f_{i,j} & \texttt{if} \quad f_{i,j} > 0 \\ 0 & \texttt{otherwise} \end{cases}$$

  – Onde o logaritmo utiliza base 2

- A expressão com logaritmo é a forma preferível porque torna os pesos diretamente comparáveis ao IDF (discutido posteriormente)

log *tf* para uma coleção de 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

Vocabulary	
1	to
2	do
3	is
4	be
5	or
6	not
7	I
8	am
9	what
10	think
11	therefore
12	da
13	let
14	it

$tf_{i,1}$	$tf_{i,2}$	$tf_{i,3}$	$tf_{i,4}$
3	2	-	-
2	-	2.585	2.585
2	-	-	-
2	2	2	2
-	1	-	-
-	1	-	-
-	2	2	-
-	2	1	-
-	1	-	-
-	-	1	-
-	-	1	-
-	-	-	2.585
-	-	-	2
-	-	-	2

$$tf_{i,j} = \begin{cases} 1 + \log f_{i,j} & \texttt{if } f_{i,j} > 0 \\ 0 & \texttt{otherwise} \end{cases}$$

Ponderação pela frequência inversa de documentos

- Chama-se **exaustividade** o número de termos de indexação associados a um documento
- Quanto maior o número de termos de indexação associados a um documento, maior é a probabilidade de recuperação daquele documento
  - Se muitos termos são associados a um documento, ele poderá ser recuperado por consultas para as quais ele não é relevante

Ponderação pela frequência inversa de documentos

- **Especificidade** é uma propriedade da semântica do termo
  - Um termo é mais ou menos específico dependendo do seu significado
  - Ex.: *bebida* é menos específico do que *chá* e *cerveja* → Pode-se esperar que *bebida* ocorra em mais documentos do que *chá* e *cerveja*
- Especificidade do termo pode ser interpretada como uma propriedade **estatística** (o inverso do número de documentos nos quais o termo ocorre)

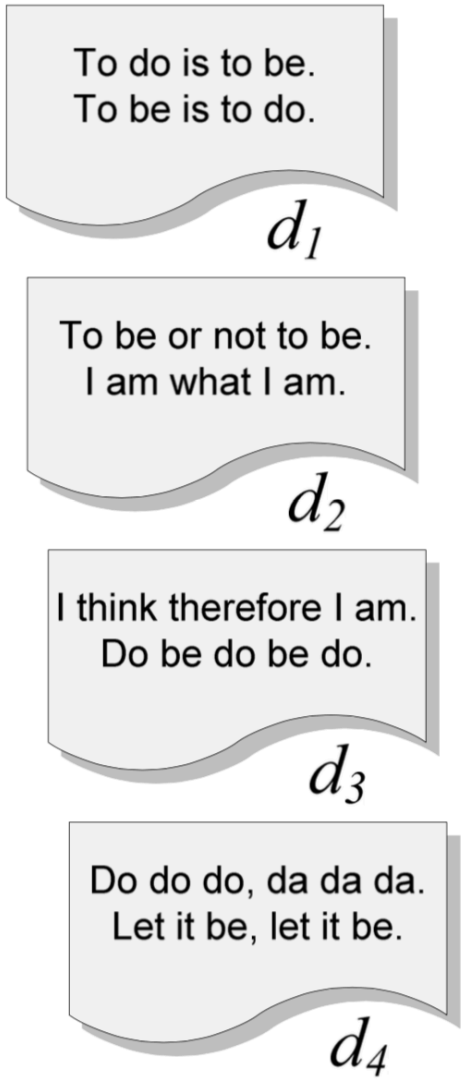
Ponderação pela frequência inversa de documentos

- Seja k_i um termo com a frequência na coleção n_i. Então,

$$idf_i = \log (N / n_i)$$

  - onde idf_i é chamado de **frequência inversa de documentos** do termo k_i
- IDF fornece a base para os esquemas de ponderação modernos e é usado por quase todos os sistemas modernos de RI

IDF para uma coleção de exemplo



	term	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 TF-IDF

- O esquema de ponderação de termos mais popular utiliza pesos que combinam os fatores IDF e as frequências dos termos
- Seja $w_{i,j}$ o peso do termo associado ao termo k_i e ao documento d_j
- Então, definimos

$$w_{i,j} = \begin{cases} (1 + \log f_{i,j}) \times \log \frac{N}{n_i} & \texttt{if} \quad f_{i,j} > 0 \\ 0 & \texttt{otherwise} \end{cases}$$

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

TF-IDF para uma coleção de 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

		d_1	d_2	d_3	d_4
1	to	3	2	-	-
2	do	0.830	-	1.073	1.073
3	is	4	-	-	-
4	be	-	-	-	-
5	or	-	2	-	-
6	not	-	2	-	-
7	I	-	2	2	-
8	am	-	2	1	-
9	what	-	2	-	-
10	think	-	-	2	-
11	therefore	-	-	2	-
12	da	-	-	-	5.170
13	let	-	-	-	4
14	it	-	-	-	4

Variantes do esquema TF-IDF

- Esquemas recomendados de ponderação TF-IDF [Salton, 1971]

Esquema de ponderação	Pesos para os termos dos docs	Pesos para os termos das consultas
1	$f_{i,j} * \log \frac{N}{n_i}$	$(0.5 + 0.5 \frac{f_{i,q}}{max_i \ f_{i,q}}) * \log \frac{N}{n_i}$
2	$1 + \log f_{i,j}$	$\log(1 + \frac{N}{n_i})$
3	$(1 + \log f_{i,j}) * \log \frac{N}{n_i}$	$(1 + \log f_{i,q}) * \log \frac{N}{n_i}$

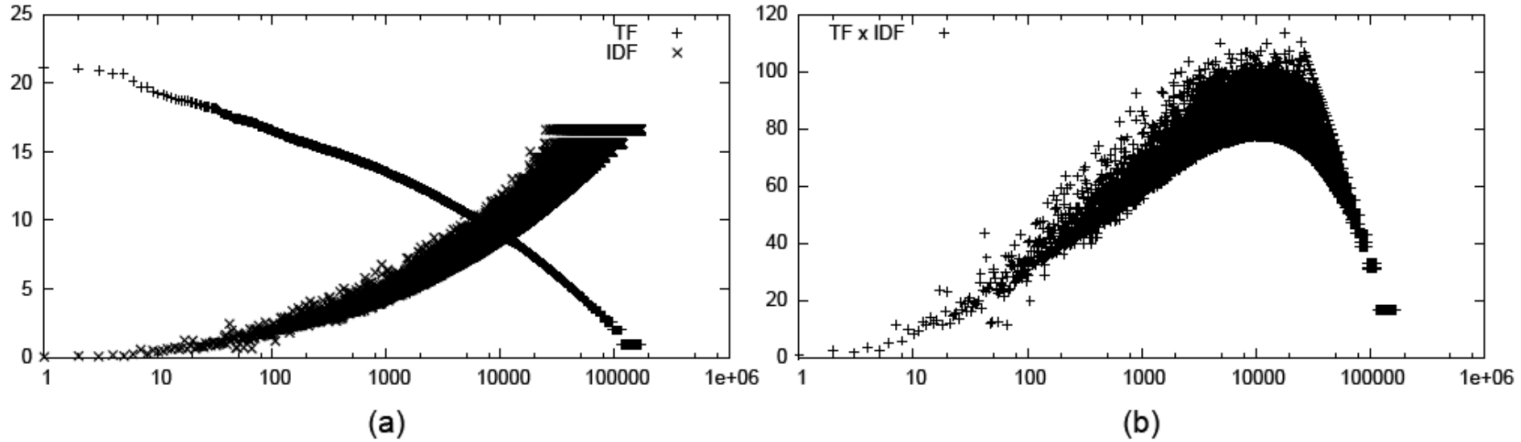
Propriedades do TF-IDF

- Considerando os pesos *tf*, *idf*, and *tf-idf* para a coleção de referência *Wall Street Journal*
- Para estudar o comportamento dos pesos, eles são plotados juntos
- Enquanto *idf* é computado sobre toda a coleção, *tf* é computado em uma base por documento. Então, precisamos de uma representação de *tf* baseada em toda a coleção, que é provida pela frequência de termos na coleção F_i
- Este raciocínio leva aos seguintes esquemas de ponderação *tf* e *idf*:

$$tf_i = 1 + \log \sum_{j=1}^{N} f_{i,j} \qquad idf_i = \log \frac{N}{n_i}$$

Propriedades do TF-IDF

- Plotando *tf* e *idf* em escala logarítmica



- Observamos que os pesos *tf* e *idf* apresentam um comportamento de lei de potência que equilibram um ao outro
- Termos com valores intermediários de *idf* atigem os valores máximos de *tf-idf* e são mais importantes para fins de ranqueamento

*Plotados em ordem decrescente *tf*

Exercício - Ponderação de termos

- Partir da implementação desenvolvida na aula anterior (modelo booleano), sendo que foram aplicadas as mesmas etapas de pré-processamento indicadas previamente.
- Implemente a ponderação TF-IDF para atribuição dos pesos dos termos (você deve utilizar o 3º esquema de ponderação sugerido por [Salton, 1971] – slide 18).

Exercício - Relembrando

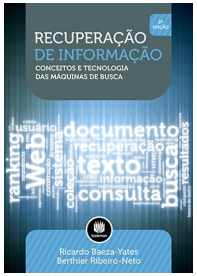
- Exemplo de entradas:

```
M=['O peã e o caval são pec de xadrez. O caval é
o melhor do jog.';
'A jog envolv a torr, o peã e o rei.';
'O peã lac o boi';
'Caval de rodei!';
'Polic o jog no xadrez.']; //conjunto de
documentos
stopwords=['a', 'o', 'e', 'é', 'de', 'do', 'no',
'são']; //lista de stopwords
q='xadrez peã caval torr'; //consulta
separadores=[' ',',',',','.','!','?']; //separadores
para tokenizacao
```

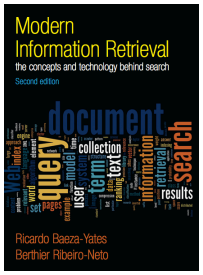
Exercício - Relembrando

- Sua implementação deve:
  - Tokenizar os documentos utilizando os separadores adequados
  - Normalizar termos (ex. caixa-baixa) e eliminar stopwords das consultas e documentos
  - Usar uma solução de indexação utilizando uma variação da matriz de incidências (obs.: guarde a frequência de aparecimento dos termos em cada documento)

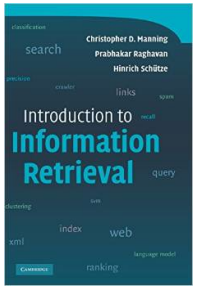
Referências



- Baeza-Yates, R.; Ribeiro-Neto, B. Recuperação de Informação: Conceitos e Tecnologia das Máquinas de Busca. 2 ed. Bookman, 2013.



- Baeza-Yates, R.; Ribeiro-Neto, B. Modern Information Retrieval. Wokingham, UK: Addison-Wesley, 2 ed., 2011.



- Manning, C. D.; Raghavan, P.; Schütze, H. Introduction to Information Retrieval. Cambridge University Press, 2008.

  Online edition 2009: http://nlp.stanford.edu/IR-book/

Referências

- Salton, G. The SMART Retrieval System – Experiments in Automatic Document Processing. Prentice-Hall Inc., Englewood Cliffs, New Jersey, 1971.

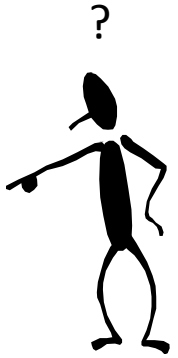
Universidade Federal do Rio de Janeiro (UFRJ)
Departamento de Ciência da Computação (DCC)

DEPARTAMENTO DE CIÊNCIA DA COMPUTAÇÃO - UFRJ -

Recuperação da Informação (MAB605)

Dúvidas?

Profa. Giseli Rabello Lopes
giseli@dcc.ufrj.br
CCMN - DCC - Sala E-2012



Material selecionado e traduzido de slides do capítulo 2 do livro [Baeza-Yates & Ribeiro-Neto, 2013]