GSI024 - Organização e recuperação de informação

Prof. Dr. Rodrigo Sanches Miani (FACOM/UFU)

Última atualização - Maio/2022

QP1

QP-1

- Dia 26/05, no horário da aula;
- Prova no formato de quiz (testes curtos - avaliar a compreensão do conteúdo);
- A prova será feita usando o computador - via formulário do Teams (fiquem à vontade para trazer o próprio computador);
- Assunto: tópicos 1 e 2.
    - Deixei uma lista de exercícios no Teams. Vários exercícios já podem ser feitos!

Ponderação de termos

Agenda

“Ponderação de termos”

Breve discussão sobre ponderação de termos

Ponderação TF

Ponderação IDF

Ponderação TF-IDF

Algumas propriedades do TF-IDF

Variantes e melhorias

Aula passada

Modelo booleano - Definição

No modelo Booleano, uma consulta q é uma expressão Booleana convencional sobre termos de indexação. Considere c(q) como qualquer dos componentes conjuntivos da consulta. Dado um documento d_j, sendo $c(d_j)$ seu componente conjuntivo de documento correspondente, então a similaridade entre o documento e a consulta q é definida por:

$$sim(d_j, q) = \begin{cases} 1 & \text{se } \exists c(q) \mid c(q) = c(d_j) \\ 0 & \text{caso contrário} \end{cases}$$

Se sim(d_j, q) = 1, então d_j é relevante a consulta q.

Modelo booleano - Exemplo 2

- O que aconteceria se a consulta fosse “Brutus AND Casear AND NOT Calpurnia”, ou seja, quais documentos satisfazem essa consulta?

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	1	1	0	0	0	1
Brutus	1	1	0	1	0	0
Caesar	1	1	0	1	1	1
Calpurnia	0	1	0	0	0	0
Cleopatra	1	0	0	0	0	0
mercy	1	0	1	1	1	1
worser	1	0	1	1	1	0

Modelo booleano – Vantagens e desvantagens

- Vantagens
  - Formalismo claro;
  - Simplicidade;
  - Fácil de implementar;
  - Adoção de pesos binários para os termos de indexação.
- Desvantagens
  - Impossibilidade de realizar ranqueamento dos documentos;
  - Formulação de consultas booleanas pode ser incoveniente para os usuários.

Breve discussão sobre ponderação de termos

Relevância de termos

- Dado um conjunto de termos de indexação para um documento, notamos que **nem todos os termos** são igualmente úteis para descrever o conteúdo dos documentos;

- De fato, existem termos de indexação que são mais vagos do que outros (Exemplo: palavras chave muito genéricas em artigos);

- Decidir a importância de um termo na sumarização do conteúdo de um documento não é uma tarefa trivial.

Relevância de termos

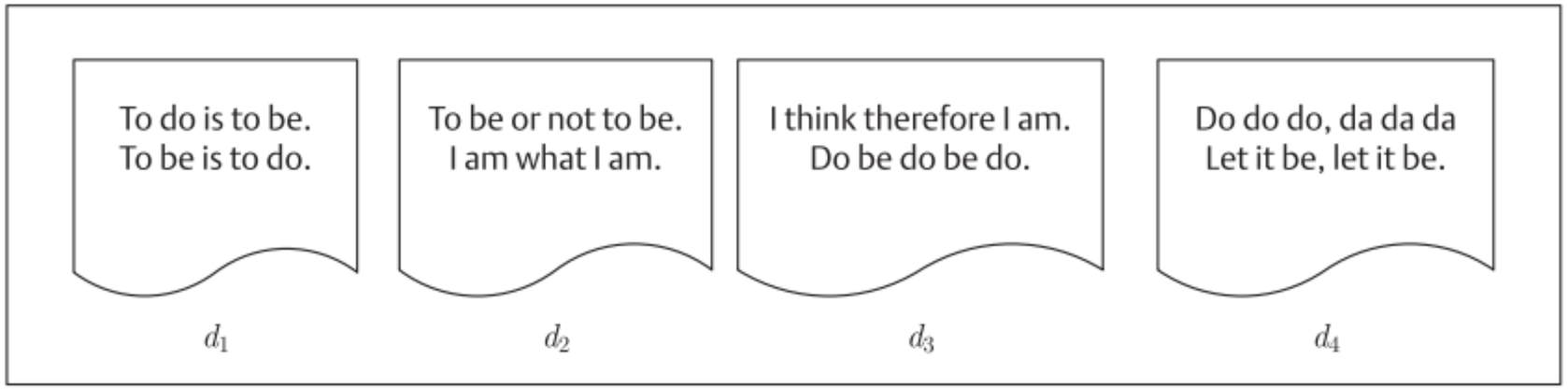
Apesar dessa dificuldade, existem propriedades de um termo de indexação que são facilmente mensuráveis e que são úteis na avaliação da importância de um termo.

Exemplo:

Considere uma coleção com cem mil documentos. Uma palavra que aparece em todos esses cem mil documentos não é útil como termo de indexação. Por outro lado, uma palavra que aparece em apenas cinco documentos dessa coleção é bastante útil, porque ela reduz consideravelmente o conjunto de documentos que podem ser do interesse do usuário.

Relevância de termos

Diferentes termos de indexação têm diferentes graus de importância para fins de descrição do conteúdo dos documentos.



Relevância de termos

- Diferentes termos de indexação têm diferentes graus de importância para fins de descrição do conteúdo dos documentos.

- Como capturar esse efeito?
  - Resposta: atribuindo pesos numéricos a cada termo de indexação de um documento.

Ponderação de termos

Definição:

A fim de caracterizar a importância dos termos, um peso $w_{i,j}$, com $w_{i,j} > 0$, é associado a cada termo de indexação k_i de um documento d_j na coleção. Para um termo de indexação k_i que não aparece no documento, $w_{i,j} = 0$.

Ponderação de termos

- O peso quantifica a importância do termo de indexação na descrição do conteúdo do documento;
- Pela atribuição de pesos aos termos de indexação conseguimos computar um **grau numérico para cada documento da coleção** em relação à consulta dada, e isso propicia melhores resultados;
- Para gerar os pesos que serão atribuídos aos termos de indexação, consideramos *o quão importante o termo é para descrever* ***um documento*** *ou* ***um subconjunto*** *de documentos da coleção.*

Ponderação de termos - Conclusão

Os termos não são igualmente importantes e devem ser ponderados de forma diferente!

Como contabilizar a importância dos termos?

Ponderação de termos – Frequência de ocorrências

Uma técnica amplamente utilizada é o cálculo da frequência de ocorrência dos termos nos documentos:

Seja $f_{i,j}$ a frequência de ocorrência do termo de indexação k_i no documento d_j, isto é, o número de vezes que o termo k_i aparece no texto do documento d_j. A frequência total F_i do termo k_i na coleção é a soma das frequências de ocorrência do termo em todos os documentos, isto é,

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

onde N é o número de documentos na coleção.

Frequência de ocorrências - Exemplo

Considere d_1 = {duas primeiras estrofes do hino nacional brasileiro} e d_2 = {estrofes 3 e 4 do hino nacional brasileiro}. Suponha que o vocabulário dessa coleção é formado por V = {ipiranga, liberdade, patria, brasil, ceu, amada, terra, salve, luz}. Monte a matriz de termos e documentos para cada termo. Encontre a frequência total para cada termo de indexação.

Frequência de ocorrências - Exemplo

D1 = {Ouviram do Ipiranga as margens plácidas De um povo heroico o brado retumbante E o sol da liberdade, em raios fúlgidos Brilhou no céu da pátria nesse instante Se o penhor dessa igualdade Conseguimos conquistar com braço forte Em teu seio, ó liberdade Desafia o nosso peito a própria morte!}

D2 = {Ó pátria amada Idolatrada Salve! Salve! Brasil, um sonho intenso, um raio vívido De amor e de esperança à terra desce Se em teu formoso céu, risonho e límpido A imagem do cruzeiro resplandece}

Ponderação TF

Método de ponderação de termos

Com base no que já vimos sobre a importância da frequência dos termos para recuperar informação, estudaremos o método mais popular de ponderação em RI, chamado de **TF-IDF** (*Term frequency – Inverse Document Frequency*).

Ponderação TF

A primeira forma de ponderação da frequência dos termos foi proposta por Luhn (1957) e baseia-se na seguinte suposição:

Hipótese de Luhn: *O valor ou peso de um termo k_i que ocorre em um documento d_j é simplesmente proporcional à frequência do termo $f_{i,j}$. Isto é, quanto mais frequentemente um termo k_i ocorrer no texto do documento d_j maior será a sua frequência de termo $TF_{i,j}$.*

Ponderação TF

- Essa hipótese baseia-se na observação que termos com alta frequência são importantes para descrever os tópicos-chave de um documento, a qual leva diretamente à seguinte formulação da ponderação TF:

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

- Ou seja, o peso do termo é dado simplesmente pela frequência do termo no documento.

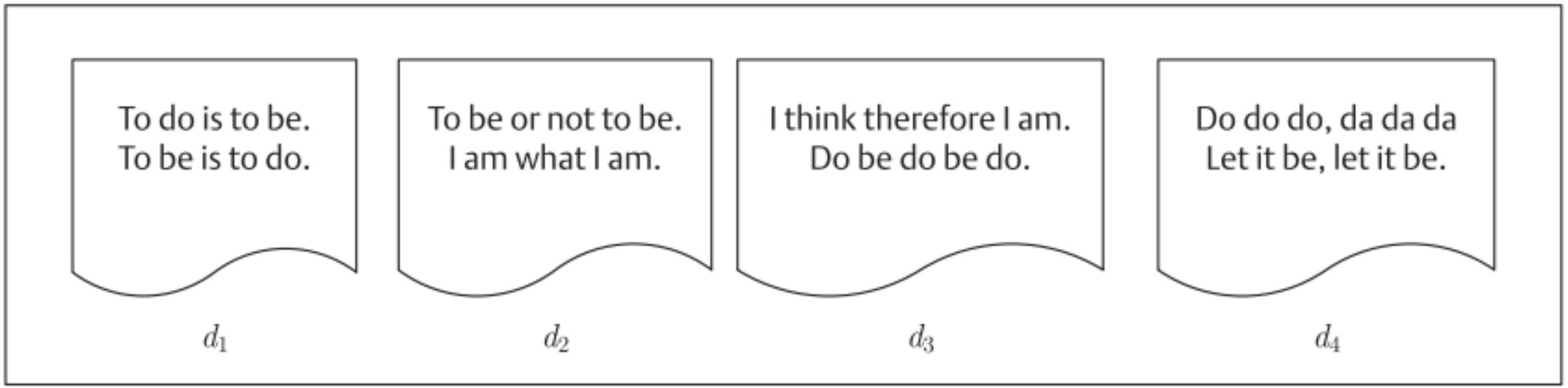
Ponderação TF - Variante

- 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 logartimo utiliza a base 2.

Ponderação TF - Exemplo



Ponderação TF - Exemplo

#	termo	$f_{i,1}$	$f_{i,2}$	$f_{i,3}$	$f_{i,4}$	$TF_{i,1}$	$TF_{i,2}$	$TF_{i,3}$	$TF_{i,4}$
1	to	4	2	–	–	3	2	–	–
2	do	2	–	3	3	2	–	2,585	2,585
3	is	2	–	–	–	2	–	–	–
4	be	2	2	2	2	2	2	2	2
5	or	–	1	–	–	–	1	–	–
6	not	–	1	–	–	–	1	–	–
7	I	–	2	2	–	–	2	2	–
8	am	–	2	1	–	–	2	1	–
9	what	–	1	–	–	–	1	–	–
10	think	–	–	1	–	–	–	1	–
11	therefore	–	–	1	–	–	–	1	–
12	da	–	–	–	3	–	–	–	2,585
13	let	–	–	–	2	–	–	–	2
14	it	–	–	–	2	–	–	–	2
Tamanho do documento (# palavras)		10	11	10	12				

Ponderação IDF

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 entender a forma com que a ponderação TF atua e como uma nova visão sobre ela pode ser feita.

Ponderação TF - Visão

- Ponderação TF analisa cada documento de maneira individual;

- Termos de indexação com maiores pesos serão aqueles que aparecem mais vezes (frequência de ocorrência) naquele documento.

Ponderação TF - Problema

Como modelar a situação em que um mesmo termo aparece em muitos documentos da **coleção**? Ou seja, como lidar com a especificidade desse termo perante toda a **coleção** de documentos?

Ideia - IDF

- Ponderação TF trata do peso dos termos perante os documentos de forma **individual**;

- Modelar a situação descrita anteriormente (termos aparecerem muitas vezes ou poucas vezes em uma coleção) exige analisar o comportamento do termo em toda a **coleção** de documentos.

Especificidade x Ponderação de termos

- Ideia: se um termo ocorrer em todos os documentos da coleção, sua ***especificidade*** é mínima e o termo não é útil para a recuperação;
- A ***especificidade*** de um termo é interpretada como o quão bem um termo descreve o tópico (assunto) de um documento;
  - Poderíamos pensar que termos com um bom grau de especificidade deveriam aparecer em poucos documentos.

Especificidade x Ponderação de termos

Problema:

Com base a ideia de especificidade, gostaríamos de um modelo de ponderação (uma fórmula!!) que fizesse o seguinte:

1. o valor do peso do termo de indexação será zero se ele puder ser encontrado em todos os documentos da coleção;

2. o valor do peso do termo aumentará se ele estiver presente em poucos documentos.

Especificidade x Ponderação de termos

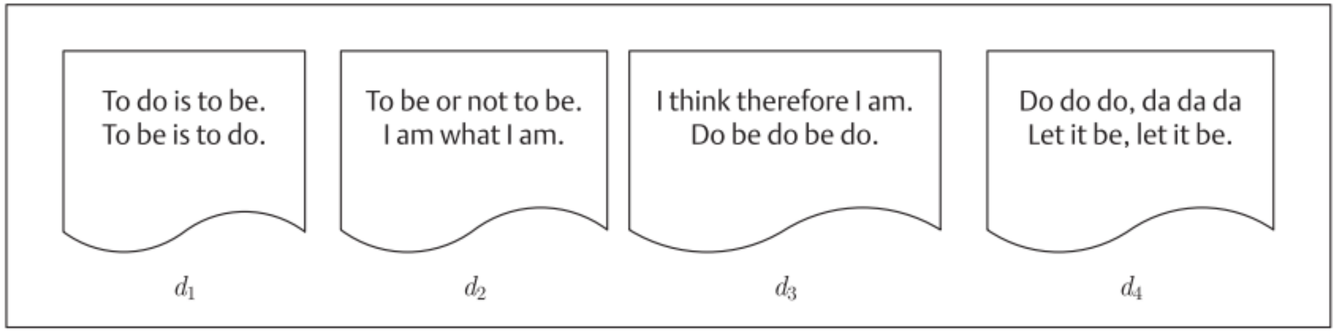
1. Verificar a ocorrência do termo k_i para cada documento d_j da coleção => n_i;
2. Calcular a frequência relativa inversa de cada termo => N/n_i;
3. Aplicar a função log (na base 2) na frequência relativa inversa de cada termo.

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

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

Ponderação IDF - Exemplo

#	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 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 atuais 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.

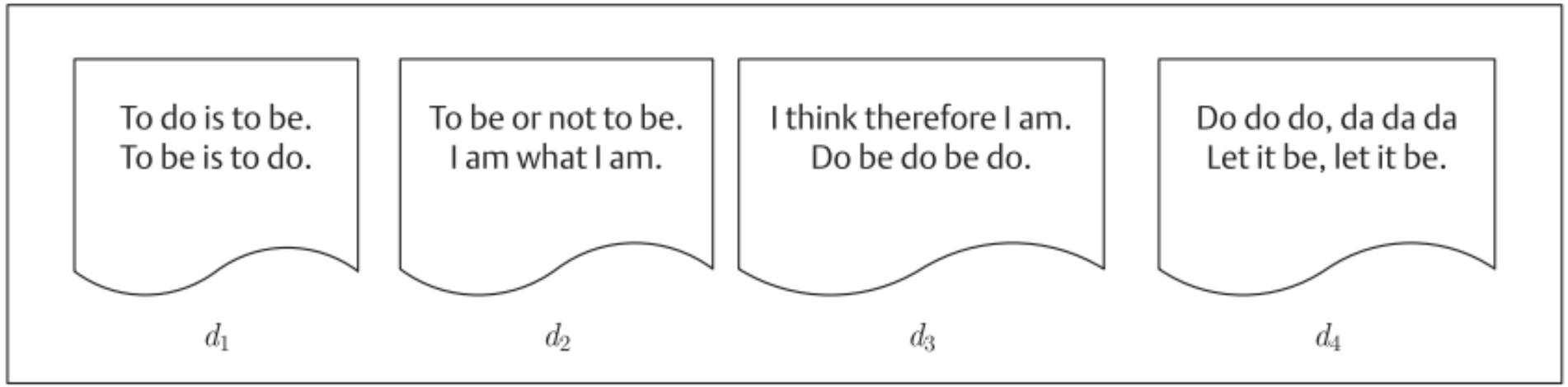
Ponderação TF-IDF - Definição

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



Ponderação TF-IDF - Exemplo

#	termo	$TF_{i,1}$	$TF_{i,2}$	$TF_{i,3}$	$TF_{i,4}$	$IDF_i = \log(N/n_i)$	d_1	d_2	d_3	d_4
1	to	3	2	–	–	1	3	2	–	–
2	do	2	–	2,585	2,585	0,415	0,830	–	1,073	1,073
3	is	2	–	–	–	2	4	–	–	–
4	be	2	2	2	2	0	–	–	–	–
5	or	–	1	–	–	2	–	2	–	–
6	not	–	1	–	–	2	–	2	–	–
7	I	–	2	2	–	1	–	2	2	–
8	am	–	2	1	–	1	–	2	1	–
9	what	–	1	–	–	2	–	2	–	–
10	think	–	–	1	–	2	–	–	2	–
11	therefore	–	–	1	–	2	–	–	2	–
12	da	–	–	–	2,585	2	–	–	–	5,170
13	let	–	–	–	2	2	–	–	–	4
14	it	–	–	–	2	2	–	–	–	4
	Tamanho do documento (normas dos vetores)						5,068	4,899	3,762	7,738

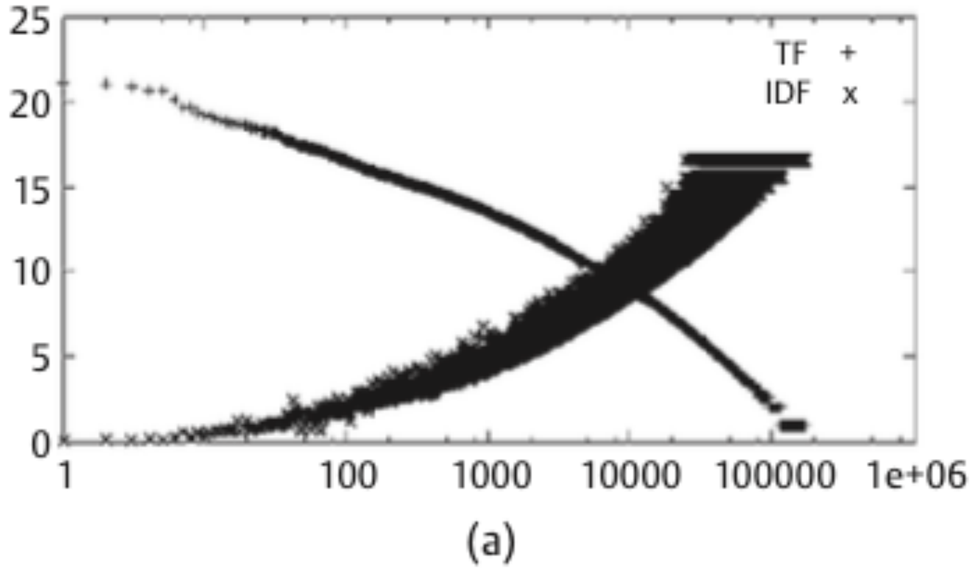
Ponderação TF-IDF – Características

- Termos mais frequentes dentro de um documento e termos mais raros dentro da coleção possuem um peso TF-IDF maior;

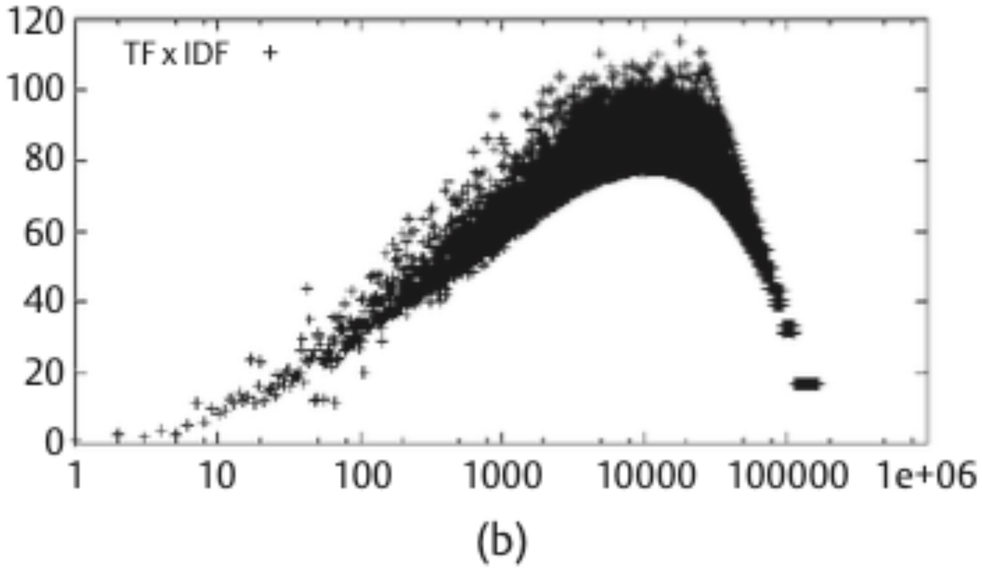
- Embora simples, os pesos TF-IDF são bastante eficazes, especialmente, para coleções genéricas:
  - Coleção de documentos sobre a qual não temos nenhuma informação.

Algumas propriedades do TF-IDF

Ponderação TF-IDF – Propriedades



(a)



(b)

Variantes e melhorias

Ponderação TF-IDF – Variantes

Esquema de ponderação	Peso para os termos dos documentos	Peso para os termos das consultas
1	$f_{i,j} \times \log \frac{N}{n_i}$	$(0{,}5 + 0{,}5 \frac{f_{i,q}}{max_i \ f_{i,q}}) \times \log \frac{N}{n_i}$
2	$1 + \log f_{i,j}$	$\log(1 + \frac{N}{n_i})$
3	$(1 + \log f_{i,j}) \times \log \frac{N}{n_i}$	$(1 + \log f_{i,q}) \times \log \frac{N}{n_i}$

Melhorias

- **Normalização** pelo tamanho dos documentos (usar a norma do vetor, por exemplo);
- **Correlação** entre termos - as ocorrências dos termos de indexação podem estar correlacionadas. Por exemplo, os termos *computador* e *rede* em uma coleção sobre redes de computadores podem “atrair" a ocorrência do outro. É possível modelar essa correlação (páginas 33 e 34 do livro);
  - Assumir a independência dos termos simplifica o processo de geração do vocabulário e, consequentemente, do ranking;
  - Não existe consenso sobre o quão útil é o uso da correlação de termos para modelar coleções genéricas, portanto, assumiremos ao longo da disciplina a independência dos termos.

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.

Próximas aulas

- Modelo vetorial;
- Modelo probabilístico;
- Mais laboratórios!!

Estudos

- Resolver a Lista 1
- Recuperação de Informação: Conceitos e Tecnologia das Máquinas de Busca
  - Capítulo 2.2.3, 2.2.4 e 2.2.5