INSTITUTO FEDERAL
MINAS GERAIS

RECUPERAÇÃO DE INFORMAÇÃO

PROFA. 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 BOLEANO ... CONTINUANDO

ROTEIRO

- Ponderação de Termos;
- Ponderação TF;

BREVE RESUMO DA AULA ANTERIOR!

MODELO BOOLEANO

- O modelo Booleano prevê que cada documento seja **relevante** ou **não relevante**;
  - Não existe satisfação parcial das condições da consulta;

- Esse critério binário de decisão, sem nenhuma noção de grau, impede uma boa qualidade na recuperação de informação.

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;
  - 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 indexação 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.

- **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

- 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;

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.

EXEMPLO

- Considere d1 = {primeira parte do hino nacional brasileiro} e d2 = {segunda parte do hino nacional brasileiro}.
  - Os termos de indexação k = {ipiranga, patria, brasil, terra}.

- Monte a matriz de termos e documentos para cada termo. Encontre a frequência total para cada termo de indexação.

- Ouviram do Ipiranga as margens plácidas
  De um povo heróico 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! 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. Gigante pela própria natureza,
  És belo, és forte, impávido colosso,
  E o teu futuro espelha essa grandeza. Terra adorada,
  Entre outras mil,
  És tu, Brasil,
  Ó Pátria amada!
  Dos filhos deste solo és mãe gentil,
  Pátria amada,
  Brasil!

- Ouviram do Ipiranga as margens plácidas
  De um povo heróico 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! 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. Gigante pela própria natureza,
  És belo, és forte, impávido colosso,
  E o teu futuro espelha essa grandeza. Terra adorada,
  Entre outras mil,
  És tu, Brasil,
  Ó Pátria amada!
  Dos filhos deste solo és mãe gentil,
  Pátria amada,
  Brasil!

TERMO	fi,1	fi,2
ipiranga	1	
patria	4	
brasil	3	
terra	2	

- Deitado eternamente em berço esplêndido,
  Ao som do mar e à luz do céu profundo,
  Fulguras, ó Brasil, florão da América,
  Iluminado ao sol do Novo Mundo! Do que a terra, mais garrida,
  Teus risonhos, lindos campos têm mais flores;
  "Nossos bosques têm mais vida",
  "Nossa vida" no teu seio "mais amores." Ó Pátria amada,
  Idolatrada,
  Salve! Salve! Brasil, de amor eterno seja símbolo
  O lábaro que ostentas estrelado,
  E diga o verde-louro dessa flâmula
  - "Paz no futuro e glória no passado."
  Mas, se ergues da justiça a clava forte,
  Verás que um filho teu não foge à luta,
  Nem teme, quem te adora, a própria morte. Terra adorada,
  Entre outras mil,
  És tu, Brasil,
  Ó Pátria amada!
  Dos filhos deste solo és mãe gentil,
  Pátria amada,
  Brasil!

- Deitado eternamente em berço esplêndido,
  Ao som do mar e à luz do céu profundo,
  Fulguras, ó Brasil, florão da América,
  Iluminado ao sol do Novo Mundo! Do que a terra, mais garrida,
  Teus risonhos, lindos campos têm mais flores;
  "Nossos bosques têm mais vida",
  "Nossa vida" no teu seio "mais amores." Ó Pátria amada,
  Idolatrada,
  Salve! Salve! Brasil, de amor eterno seja símbolo
  O lábaro que ostentas estrelado,
  E diga o verde-louro dessa flâmula
  - "Paz no futuro e glória no passado."
  Mas, se ergues da justiça a clava forte,
  Verás que um filho teu não foge à luta,
  Nem teme, quem te adora, a própria morte. Terra adorada,
  Entre outras mil,
  És tu, Brasil,
  Ó Pátria amada!
  Dos filhos deste solo és mãe gentil,
  Pátria amada,
  Brasil!

TERMO	fi,1	fi,2
ipiranga	1	0
patria	4	3
brasil	3	4
terra	2	2

PONDERAÇÃO TF

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:

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

- onde o logaritmo utiliza a base 2

EXEMPLO

TERMO	fi,1	fi,2
ipiranga	1	0
patria	4	3
brasil	3	4
terra	2	2

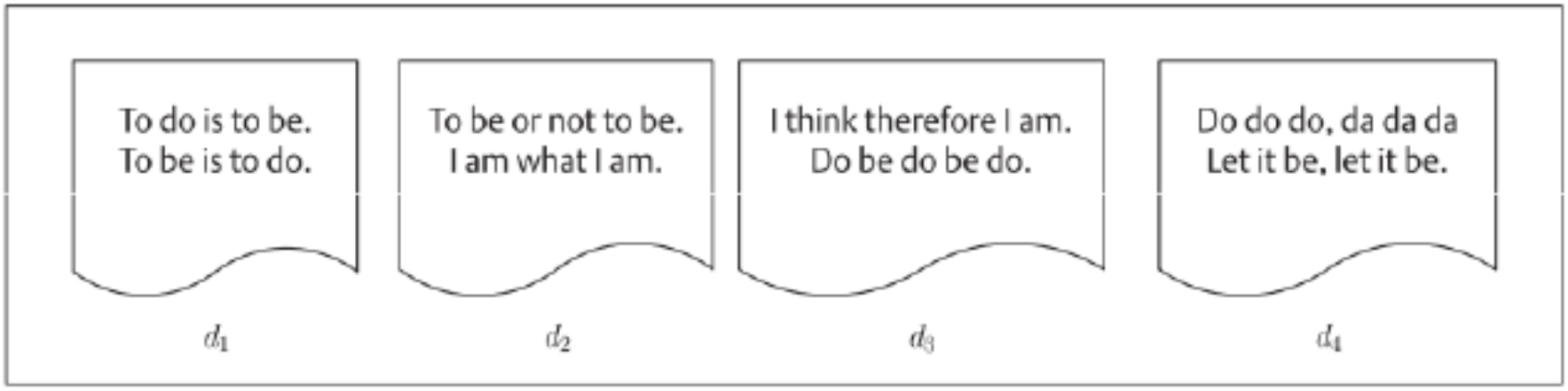
$$tf_{i,j} = \begin{cases} 1 + \log f_{i,j} & se\ f_{i,j} > 0 \\ 0 & caso\ contrário \end{cases}$$

TERMO	Log	Log
ipiranga	0	
patria	2	1,584963
brasil	1,584963	2
terra	1	1

TERMO	tfi,1	tfi,2
ipiranga	1	0
patria	3	2,584963
brasil	2,584963	3
terra	2	2

ATIVIDADE PARA ENTREGAR ATÉ DIA 30/05

- Considere a coleção abaixo e monte a matriz com as frequências e a matriz com o peso TF conforme a formula do slide 22. Considere como vocabulário, todos os termos.



“Dar o melhor de si é mais importante que ser o melhor”.

Mike Lermer