

# **Modelo Vetorial**

**Wendel Melo**

Faculdade de Computação  
Universidade Federal de Uberlândia

Recuperação da Informação

Adaptado do Material da Prof<sup>a</sup> Vanessa Braganholo - IC/UFF

# Modelo Vetorial

- Proposto ao final da década de 1960 e ainda hoje é um dos modelos mais empregados;
- Busca suprir as limitações do modelo booleano:
  - Casamento entre consultas e documentos precisa ser exato;
  - Falta de ranqueamento.
- Permite casamentos parciais através da adoção de pesos não binários aos termos de indexação, tanto para documentos quanto para consultas;

# Modelo Vetorial

- Documentos e consultas são representados como vetores em um espaço vetorial
  - A dimensão do espaço é dada pelo número de termos  $T$  no vocabulário da base de documentos.
- Assim, cada documento  $d_j$  é representado pelo vetor:

$$\vec{d}_j = (w_{1j}, w_{2j}, \dots, w_{Tj})$$

- E uma consulta  $q$  é representada por:

$$\vec{q} = (w_{1q}, w_{2q}, \dots, w_{Tq})$$

# Modelo Vetorial

- Documentos e consultas são representados como vetores em um espaço vetorial
  - A dimensão do espaço é dada pelo número de termos  $T$  no vocabulário da base de documentos.
- Assim, cada documento  $d_j$  é representado pelo vetor:

$$\vec{d}_j = (w_{1j}, w_{2j}, \dots, w_{Tj})$$

Os elementos  $w_{ij}$  e  $w_{iq}$  dos vetores são pesos para os termos, tanto nos docs quanto nas consultas!

- E uma consulta  $q$  é representada por:

$$\vec{q} = (w_{1q}, w_{2q}, \dots, w_{Tq})$$

# Modelo Vetorial

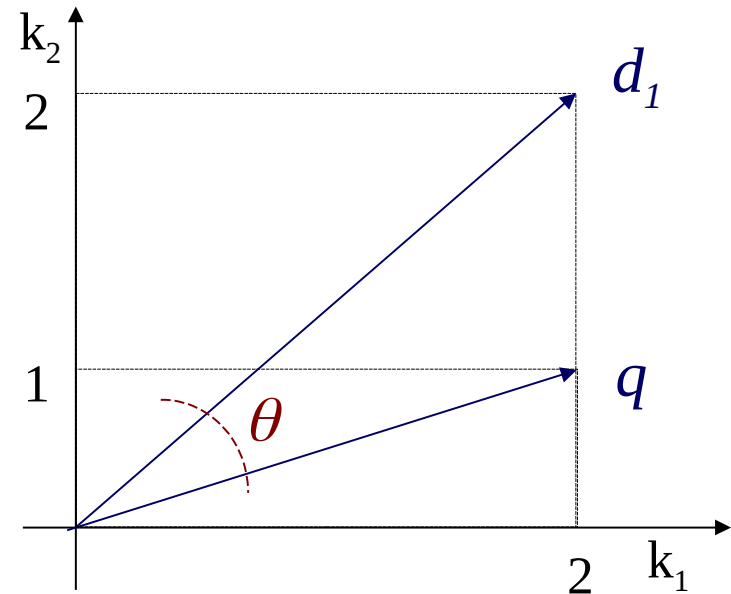
- Algum esquema de ponderação deve ser utilizado para o cálculo dos pesos  $w_{ij}$  e  $w_{iq}$ ;
  - **É comum o uso da ponderação TF-IDF!** Outros esquemas podem ser utilizados, todavia.
- Originalmente, estes pesos devem ser **não-negativos!**
- Documentos e consultas são representados como vetores em um espaço vetorial:
  - A similaridade entre um documento e uma consulta é calculada através da **similaridade entre seus respectivos vetores!**

# Modelo Vetorial

**Exemplo:** Um vocabulário de dois termos,  $k_1$  e  $k_2$ :

$$\vec{d}_1 = (2, 2)$$

$$\vec{q} = (2, 1)$$



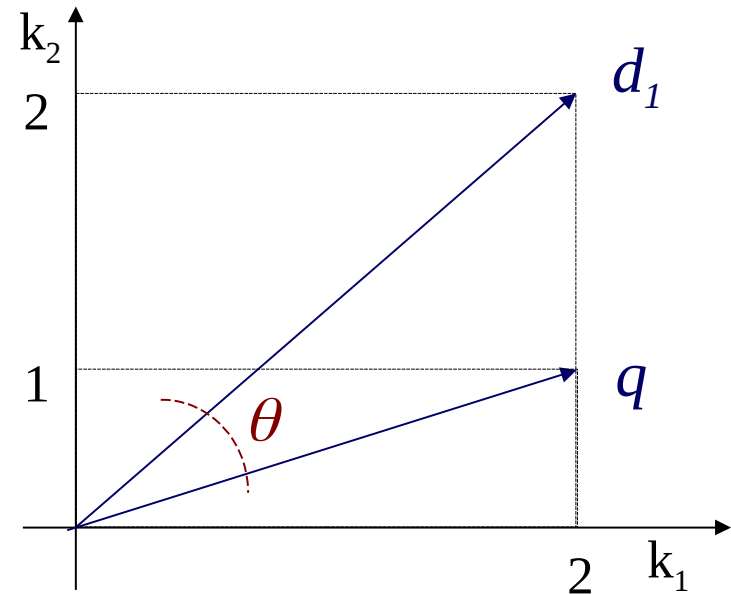
# Modelo Vetorial

**Exemplo:** Um vocabulário de dois termos,  $k_1$  e  $k_2$ :

$$\vec{d}_1 = (2, 2)$$

$$\vec{q} = (2, 1)$$

Pelo fato de haverem dois termos, os vetores estarão em  $\mathbb{R}^2$ .



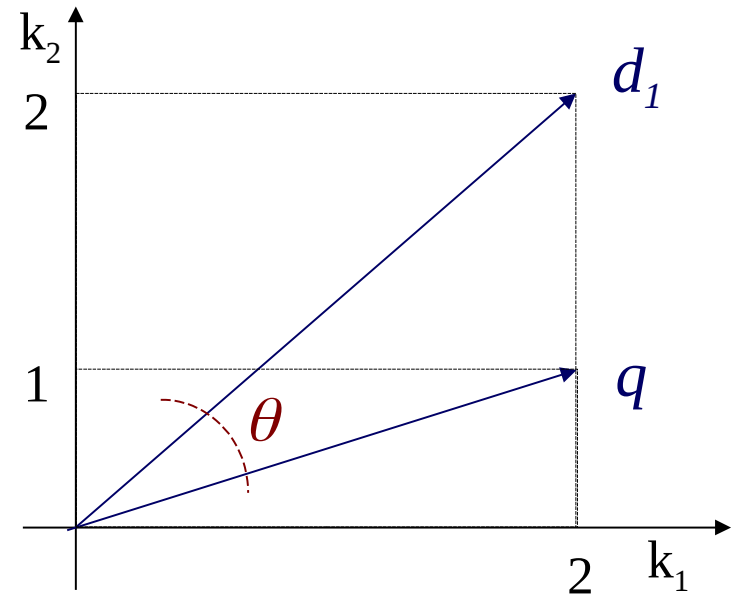
# Modelo Vetorial

**Exemplo:** Um vocabulário de dois termos,  $k_1$  e  $k_2$ :

$$\vec{d}_1 = (2, 2)$$

$$\vec{q} = (2, 1)$$

Como os pesos dos termos devem ser não negativos, os vetores dos documentos e consultas estarão no ortante positivo do espaço (no caso de  $\mathbb{R}^2$ , será o primeiro quadrante).





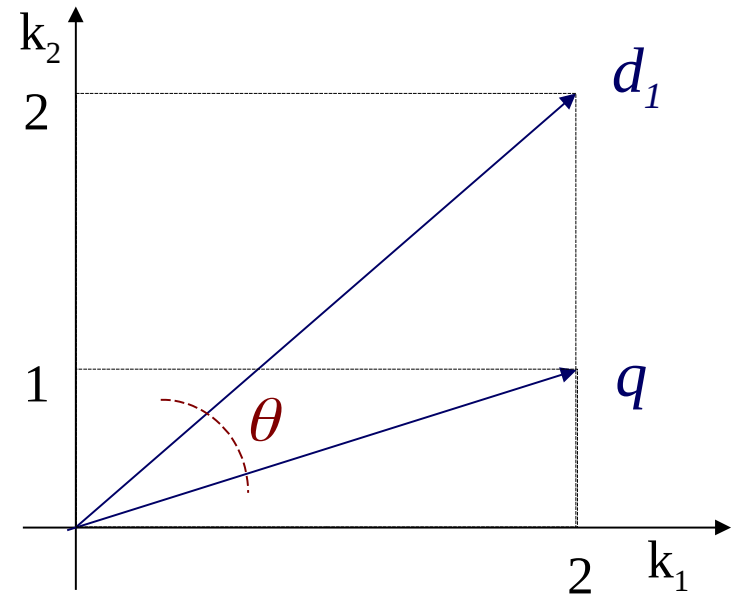
# Modelo Vetorial

**Exemplo:** Um vocabulário de dois termos,  $k_1$  e  $k_2$ :

$$\vec{d}_1 = (2, 2)$$

$$\vec{q} = (2, 1)$$

Quanto mais próximo de se sobreporem estiverem os vetores de  $d_1$  e  $q$ , maior a similaridade entre documento e consulta.



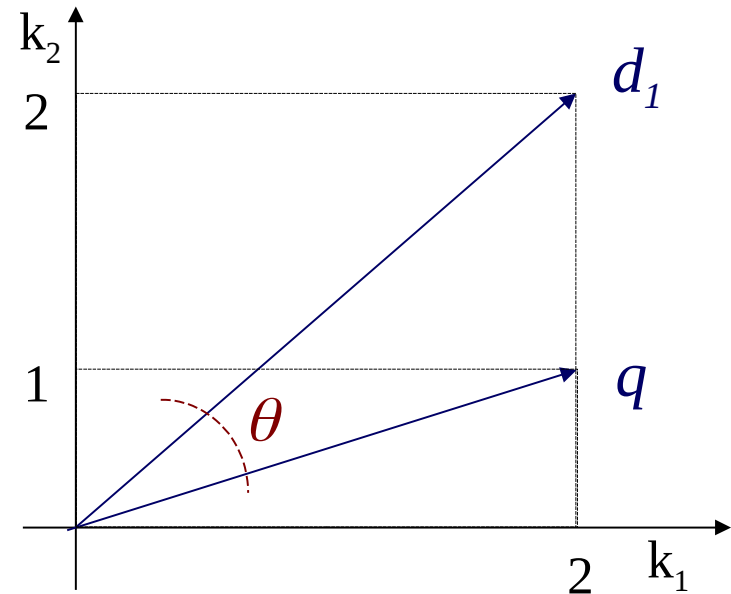
# Modelo Vetorial

**Exemplo:** Um vocabulário de dois termos,  $k_1$  e  $k_2$ :

$$\vec{d}_1 = (2, 2)$$

$$\vec{q} = (2, 1)$$

Assim, essa similaridade pode ser medida através do ângulo  $\theta$  entre os vetores. Quanto mais fechado o ângulo, maior a similaridade.



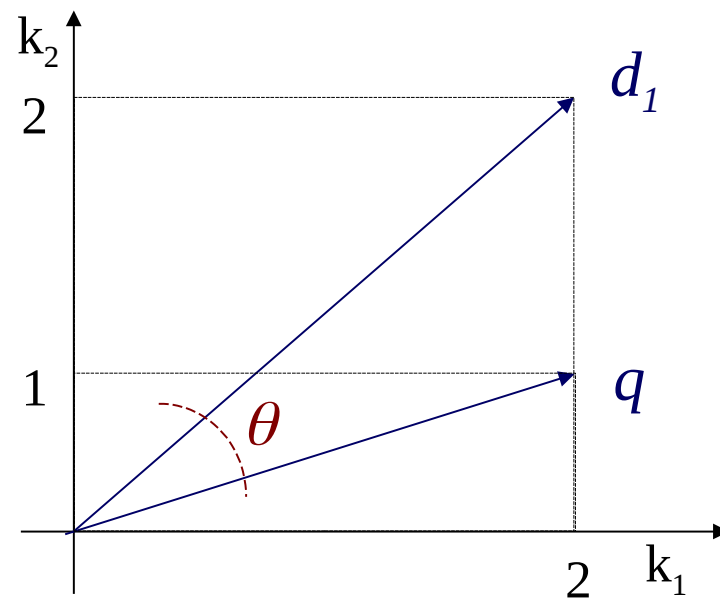
# Modelo Vetorial

**Exemplo:** Um vocabulário de dois termos,  $k_1$  e  $k_2$ :

$$\vec{d}_1 = (2, 2)$$

$$\vec{q} = (2, 1)$$

A similaridade entre o documento e a consulta é calculada através do cosseno do ângulo  $\theta$ .



# Modelo Vetorial

Dado um doc  $d_j$  representado pelo vetor:  $d_j = (w_{1j}, w_{2j}, \dots, w_{Tj})$ ,

e uma consulta  $q$  representada pelo vetor:  $q = (w_{1q}, w_{2q}, \dots, w_{Tq})$ ,

a similaridade entre o documento  $d_j$  e a consulta  $q$  é obtida através do cosseno do ângulo entre seus vetores:

$$\text{sim}(d_j, q) = \cos \theta = \frac{\sum_{i=1}^T w_{ij} \times w_{iq}}{\left( \sqrt{\sum_{i=1}^T w_{ij}^2} \right) \times \left( \sqrt{\sum_{i=1}^T w_{iq}^2} \right)}$$

, onde  $T$  é o número de termos no vocabulário

# Modelo Vetorial

- Por que é utilizado o cosseno do ângulo  $\theta$  entre os vetores?

$$\text{sim}(d_j, q) = \cos \theta = \frac{\sum_{i=1}^T w_{ij} \times w_{iq}}{\left( \sqrt{\sum_{i=1}^T w_{ij}^2} \right) \times \left( \sqrt{\sum_{i=1}^T w_{iq}^2} \right)}$$

# Modelo Vetorial

- **Por que é utilizado o cosseno do ângulo  $\theta$  entre os vetores?**
  - Porque sabemos que  $\theta$  está  $0^\circ$  e  $90^\circ$ . Assim, seu cosseno estará entre 0 e 1. Quanto menor  $\theta$ , maior seu cosseno!

$$\text{sim}(d_j, q) = \cos \theta = \frac{\sum_{i=1}^T w_{ij} \times w_{iq}}{\left( \sqrt{\sum_{i=1}^T w_{ij}^2} \right) \times \left( \sqrt{\sum_{i=1}^T w_{iq}^2} \right)}$$

# Modelo Vetorial

- **Por que é utilizado o cosseno do ângulo  $\theta$  entre os vetores?**
  - Porque sabemos que  $\theta$  está  $0^\circ$  e  $90^\circ$ . Assim, seu cosseno estará entre 0 e 1. Quanto menor  $\theta$ , maior seu cosseno!
- **Por que sabemos que  $\theta$  está  $0^\circ$  e  $90^\circ$ ?**

$$\text{sim}(d_j, q) = \cos \theta = \frac{\sum_{i=1}^T w_{ij} \times w_{iq}}{\left( \sqrt{\sum_{i=1}^T w_{ij}^2} \right) \times \left( \sqrt{\sum_{i=1}^T w_{iq}^2} \right)}$$

# Modelo Vetorial

- **Por que é utilizado o cosseno do ângulo  $\theta$  entre os vetores?**
  - Porque sabemos que  $\theta$  está  $0^\circ$  e  $90^\circ$ . Assim, seu cosseno estará entre 0 e 1. Quanto menor  $\theta$ , maior seu cosseno!
- **Por que sabemos que  $\theta$  está  $0^\circ$  e  $90^\circ$ ?**
  - Porque os pesos são não-negativos, o que faz que os vetores estejam no ortante positivo (no caso de  $R^2$ , o primeiro quadrante).

$$\text{sim}(d_j, q) = \cos \theta = \frac{\sum_{i=1}^T w_{ij} \times w_{iq}}{\left( \sqrt{\sum_{i=1}^T w_{ij}^2} \right) \times \left( \sqrt{\sum_{i=1}^T w_{iq}^2} \right)}$$



# Modelo Vetorial

- Pode-se considerar que um documento  $d_j$  é relevante para uma consulta  $q$  se a similaridade entre  $d_j$  e  $q$  é superior a um determinado **patamar mínimo de similaridade**.
- Pode-se utilizar o próprio valor da similaridade para **ranquear** os documentos!

$$\text{sim}(d_j, q) = \cos \theta = \frac{\sum_{i=1}^T w_{ij} \times w_{iq}}{\left( \sqrt{\sum_{i=1}^T w_{ij}^2} \right) \times \left( \sqrt{\sum_{i=1}^T w_{iq}^2} \right)}$$

# Exemplo - Modelo Vetorial - TF-IDF

- Considere uma base de 4 documentos com os termos A, B e C e uma consulta AB:

Documento	Conteúdo
1	AAAB
2	AAC
3	AA
4	BB

# Exemplo - Modelo Vetorial - TF-IDF

- Considere uma base de 4 documentos com os termos A, B e C e uma consulta AB:

Documento	Conteúdo
1	A A A B
2	A A C
3	A A
4	B B

Começamos pelo cálculo dos IDF's, pois não variam de acordo com o documento

# Exemplo - Modelo Vetorial - TF-IDF

- Considere uma base de 4 documentos com os termos A, B e C e uma consulta AB:

Documento	Conteúdo
1	AAAB
2	AAC
3	AA
4	BB

$N$ : número de docs ;

$f_{ij}$ : frequência do termo  $k_i$  no doc  $d_j$ ;

$n_i$ : número de docs com o termo  $k_i$ .

$$idf(A) = \log\left(\frac{N}{n_t}\right) = \log\left(\frac{4}{3}\right) = 0.1249$$

$$idf(B) = \log\left(\frac{4}{2}\right) = 0.3010$$

$$idf(C) = \log\left(\frac{4}{1}\right) = 0.6021$$

# Exemplo - Modelo Vetorial - TF-IDF

- Considere uma base de 4 documentos com os termos A, B e C e uma consulta AB:

Documento	Conteúdo
1	A A A B
2	A A C
3	A A
4	B B

Calculamos os pesos no documentos segundo a expressão TF-IDF:

$N$ : número de docs ;  
 $f_{ij}$ : frequência do termo  $k_i$  no doc  $d_j$ ;  
 $n_i$ : número de docs com o termo  $k_i$ .

$$w_{ij} = \begin{cases} (1 + \log(f_{ij})) \times idf(k_i) & , \text{ se } f_{ij} \geq 1 \\ 0 & , \text{ caso contrário} \end{cases}$$

# Exemplo - Modelo Vetorial - TF-IDF

- Considere uma base de 4 documentos com os termos A, B e C e uma consulta AB:

Documento	Conteúdo
1	AAAB
2	AAC
3	AA
4	BB

$$w_{A1} = (1 + \log 3) * 0.1249 = 0.1845$$

$$w_{B1} = (1 + \log 1) * 0.3010 = 0.3010$$

$$w_{C1} = 0$$

$$\bar{d}_1 = (0.1845, 0.3010, 0)$$

Vetor de pesos do documento 1

# Exemplo - Modelo Vetorial - TF-IDF

- Considere uma base de 4 documentos com os termos A, B e C e uma consulta AB:

Documento	Conteúdo
1	AAAB
2	AAC
3	AA
4	BB

$$w_{A1} = (1 + \log 3) * 0.1249 = 0.1845$$

$$w_{B1} = (1 + \log 1) * 0.3010 = 0.3010$$

$$w_{C1} = 0$$

$$\bar{d}_1 = (0.1845, 0.3010, 0)$$

$$w_{A2} = (1 + \log 2) * 0.1249 = 0.1625$$

$$w_{B2} = 0$$

$$w_{C2} = (1 + \log 1) * 0.6021 = 0.6021$$

$$\bar{d}_2 = (0.1625, 0, 0.6021)$$

Vetor de pesos do documento 2



# Exemplo - Modelo Vetorial - TF-IDF

- Considere uma base de 4 documentos com os termos A, B e C e uma consulta AB:

Documento	Conteúdo
1	AAAB
2	AAC
3	AA
4	BB

$$w_{A3} = (1 + \log 2) * 0.1249 = 0.1625$$

$$w_{B3} = 0$$

$$w_{C3} = 0$$

$$\bar{d}_3 = (0.1625, 0, 0)$$

Vetor de pesos do documento 3



# Exemplo - Modelo Vetorial - TF-IDF

- Considere uma base de 4 documentos com os termos A, B e C e uma consulta AB:

Documento	Conteúdo
1	AAAB
2	AAC
3	AA
4	BB

$$w_{A3} = (1 + \log 2) * 0.1249 = 0.1625$$

$$w_{B3} = 0$$

$$w_{C3} = 0$$

$$\bar{d}_3 = (0.1625, 0, 0)$$

$$w_{A4} = 0$$

$$w_{B4} = (1 + \log 2) * 0.3010 = 0.3916$$

$$w_{C4} = 0$$

$$\bar{D}_4 = (0, 0.3916, 0)$$

Vetor de pesos do documento 4



# Exemplo - Modelo Vetorial - TF-IDF

- Considere uma base de 4 documentos com os termos A, B e C e uma consulta AB:

Documento	Conteúdo
1	AAAB
2	AAC
3	AA
4	BB

$$\bar{d}_1 = (0.1845, 0.3010, 0)$$

$$\bar{d}_2 = (0.1625, 0, 0.6021)$$

$$\bar{d}_3 = (0.1625, 0, 0)$$

$$\bar{d}_4 = (0, 0.3916, 0)$$

Calculamos o vetor de pesos da consulta segundo TF-IDF:

$$w_{Aq} = (1 + \log 1) * 0.1249 = 0.1249$$

$$w_{Bq} = (1 + \log 1) * 0.3010 = 0.3010$$

$$w_{Cq} = 0$$

$$\bar{d}_q = (0.1249, 0.3010, 0)$$

# Exemplo - Modelo Vetorial - TF-IDF

- Considere uma base de 4 documentos com os termos A, B e C e uma consulta AB:

Temos os seguintes vetores representando os componentes do sistema:

$$\bar{d}_1 = (0.1845, 0.3010, 0)$$

$$\bar{d}_2 = (0.1625, 0, 0.6021)$$

$$\bar{d}_3 = (0.1625, 0, 0)$$

$$\bar{d}_4 = (0, 0.3916, 0)$$

$$\bar{d}_q = (0.1249, 0.3010, 0)$$

Calculamos a similaridade entre cada documento e a consulta segundo a expressão:

$$\text{sim}(d_j, q) = \frac{\sum_{i=1}^T w_{ij} \times w_{iq}}{\left( \sqrt{\sum_{i=1}^T w_{ij}^2} \right) \times \left( \sqrt{\sum_{i=1}^T w_{iq}^2} \right)}$$

# Exemplo - Modelo Vetorial - TF-IDF

- Considere uma base de 4 documentos com os termos A, B e C e uma consulta AB:

**Similaridade entre  $q$  e  $d_1$ :**

$$\bar{d}_1 = (0.1845, 0.3010, 0)$$

$$\bar{d}_q = (0.1249, 0.3010, 0)$$

$$\text{sim}(d_j, q) = \frac{\sum_{i=1}^T w_{ij} \times w_{iq}}{\left( \sqrt{\sum_{i=1}^T w_{ij}^2} \right) \times \left( \sqrt{\sum_{i=1}^T w_{iq}^2} \right)}$$

$$\text{sim}(d_1, q) = \frac{0.1845 \times 0.1249 + 0.3010 \times 0.3010 + 0 \times 0}{\left( \sqrt{0.0340 + 0.0906 + 0} \right) \times \left( \sqrt{0.0156 + 0.0906 + 0} \right)}$$

$$\text{sim}(d_1, q) = \frac{0.0230 + 0.0906}{\left( \sqrt{0.1246} \right) \times \left( \sqrt{0.1062} \right)} = \frac{0.1136}{0.3530 \times 0.3259} = \frac{0.1136}{0.1150} = 0.9878$$

$$\text{sim}(d_1, q) = 0.9878$$

# Exemplo - Modelo Vetorial - TF-IDF

- Considere uma base de 4 documentos com os termos A, B e C e uma consulta AB:

**Similaridade entre  $q$  e  $d_2$ :**

$$\bar{d}_2 = (0.1625, 0, 0.6021)$$

$$\bar{d}_q = (0.1249, 0.3010, 0)$$

$$\text{sim}(d_j, q) = \frac{\sum_{i=1}^T w_{ij} \times w_{iq}}{\left( \sqrt{\sum_{i=1}^T w_{ij}^2} \right) \times \left( \sqrt{\sum_{i=1}^T w_{iq}^2} \right)}$$

$$\text{sim}(d_2, q) = \frac{0.1625 \times 0.1249 + 0 \times 0.3010 + 0.6021 \times 0}{\left( \sqrt{0.0264 + 0 + 0.3625} \right) \times \left( \sqrt{0.0156 + 0.0906 + 0} \right)}$$

$$\text{sim}(d_2, q) = \frac{0.0203 + 0 + 0}{\left( \sqrt{0.3889} \right) \times \left( \sqrt{0.1062} \right)} = \frac{0.0203}{0.6236 \times 0.3259} = \frac{0.0203}{0.2032} = 0.0999$$

$$\text{sim}(d_2, q) = 0.0999$$

# Exemplo - Modelo Vetorial - TF-IDF

- Considere uma base de 4 documentos com os termos A, B e C e uma consulta AB:

**Similaridade entre  $q$  e  $d_3$ :**

$$\bar{d}_3 = (0.1625, 0, 0)$$

$$\bar{d}_q = (0.1249, 0.3010, 0)$$

$$\text{sim}(d_j, q) = \frac{\sum_{i=1}^T w_{ij} \times w_{iq}}{\left( \sqrt{\sum_{i=1}^T w_{ij}^2} \right) \times \left( \sqrt{\sum_{i=1}^T w_{iq}^2} \right)}$$

$$\text{sim}(d_3, q) = \frac{0.1625 \times 0.1249 + 0 \times 0.3010 + 0 \times 0}{\left( \sqrt{0.0264 + 0 + 0} \right) \times \left( \sqrt{0.0156 + 0.0906 + 0} \right)}$$

$$\text{sim}(d_3, q) = \frac{0.0203 + 0 + 0}{\left( \sqrt{0.0264} \right) \times \left( \sqrt{0.1062} \right)} = \frac{0.0203}{0.1625 \times 0.3259} = \frac{0.0203}{0.0530} = 0.3830$$

$$\text{sim}(d_3, q) = 0.3830$$

# Exemplo - Modelo Vetorial - TF-IDF

- Considere uma base de 4 documentos com os termos A, B e C e uma consulta AB:

**Similaridade entre  $q$  e  $d_4$ :**

$$\bar{d}_4 = (0, 0.3916, 0)$$

$$\bar{d}_q = (0.1249, 0.3010, 0)$$

$$\text{sim}(d_j, q) = \frac{\sum_{i=1}^T w_{ij} \times w_{iq}}{\left( \sqrt{\sum_{i=1}^T w_{ij}^2} \right) \times \left( \sqrt{\sum_{i=1}^T w_{iq}^2} \right)}$$

$$\text{sim}(d_4, q) = \frac{0 \times 0.1249 + 0.3916 \times 0.3010 + 0 \times 0}{\left( \sqrt{0 + 0.1534 + 0} \right) \times \left( \sqrt{0.0156 + 0.0906 + 0} \right)}$$

$$\text{sim}(d_4, q) = \frac{0 + 0.1179 + 0}{\left( \sqrt{0.1534} \right) \times \left( \sqrt{0.1062} \right)} = \frac{0.1179}{0.3917 \times 0.3259} = \frac{0.1179}{0.1277} = 0.9233$$

$$\text{sim}(d_4, q) = 0.9233$$

# Exemplo - Modelo Vetorial - TF-IDF

- Considere uma base de 4 documentos com os termos A, B e C e uma consulta AB:

Documento	Conteúdo
1	A A A B
2	A A C
3	A A
4	B B

**Assim, temos:**

$$\text{sim}(d_1, q) = 0.9878$$

$$\text{sim}(d_2, q) = 0.0999$$

$$\text{sim}(d_3, q) = 0.3830$$

$$\text{sim}(d_4, q) = 0.9233$$

Portanto, o ranqueamento final fica:

$$d_1, d_4, d_3 \text{ e } d_2.$$

Assumindo um patamar mínimo de similaridade de 0.1, o documento  $d_2$  seria considerado como não relevante e não entraria no resultado.



# Exemplo - Modelo Vetorial - TF-IDF

- Considere uma base de 4 documentos com os termos A, B e C e uma consulta AB:

Documento	Conteúdo
1	A A A B
2	A A C
3	A A
4	B B

**Assim, temos:**

$$\text{sim}(d_1, q) = 0.9878$$

$$\text{sim}(d_2, q) = 0.0999$$

$$\text{sim}(d_3, q) = 0.3830$$

$$\text{sim}(d_4, q) = 0.9233$$

Portanto, o ranqueamento final fica:

$$d_1, d_4, d_3 \text{ e } d_2.$$

Assumindo um patamar mínimo de similaridade de 0.1, o documento  $d_2$  seria considerado como não relevante e não entraria no resultado.

Por que a similaridade de  $d_4$  é superior a de  $d_3$ , se ambos só tem um dos termos da consulta aparecendo duas vezes?

# Exemplo - Modelo Vetorial - TF-IDF

- Considere uma base de 4 documentos com os termos A, B e C e uma consulta AB:

Documento	Conteúdo
1	A A A B
2	A A C
3	A A
4	B B

**Assim, temos:**

$$\text{sim}(d_1, q) = 0.9878$$

$$\text{sim}(d_2, q) = 0.0999$$

$$\text{sim}(d_3, q) = 0.3830$$

$$\text{sim}(d_4, q) = 0.9233$$

Portanto, o ranqueamento final fica:

$$d_1, d_4, d_3 \text{ e } d_2.$$

Assumindo um patamar mínimo de similaridade de 0.1, o documento  $d_2$  seria considerado como não relevante e não entraria no resultado.

Por que  $d_3$  é considerado relevante e  $d_2$  não?

# Exemplo - Modelo Vetorial - TF-IDF

- Considere uma base de 4 documentos com os termos A, B e C:

Documento	Conteúdo
1	A A A B
2	A A C
3	A A
4	B B

Para uma consulta  $q = AC$ , teríamos:

$$\text{sim}(d_1, q) = 0.1061$$

$$\text{sim}(d_2, q) = 0.9983$$

$$\text{sim}(d_3, q) = 0.2031$$

$$\text{sim}(d_4, q) = 0$$

# Vantagens do Modelo Vetorial

- Simples, rápido e de “fácil” implementação;
- Uso de ponderação melhora a qualidade da recuperação;
- Permite casamento parcial entre documentos e consultas;
- Possui ranqueamento;
- A normalização pelo tamanho do documento está naturalmente embutida na fórmula do cosseno.

# Desvantagens do Modelo Vetorial

- Termos de indexação são plotados como eixos:
  - São realmente independentes?
  - São realmente ortogonais?
- Como encontrar documentos que não contém um certo termo?