Métodos Matriciais e Análise de Clusters

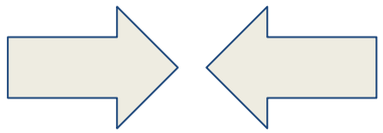
Similaridade e Distância

Laura de Oliveira Fernandes Moraes

Por que usar?

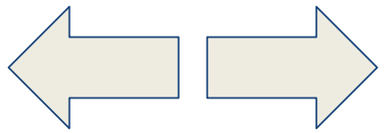
- Sistemas de recomendação (Amazon, Netflix)
- Agrupamentos
- Classificação e regressão
- Detecção de anomalias/outliers
- Descoberta de casos parecidos (diagnóstico médico, precedentes legais)

Similaridade e Distância



Similaridade

- Aumenta se os objetos são **mais** parecidos.
- Normalmente entre **0 e 1**, onde 1 representa objetos idênticos.



Distância

- Aumenta se os objetos são **menos** parecidos.
- Valor sempre será **maior ou igual a 0**, onde 0 representa objetos idênticos.

Similaridade e Distância

- Se dois objetos são representados por vetores, é possível calcular a distância entre eles:

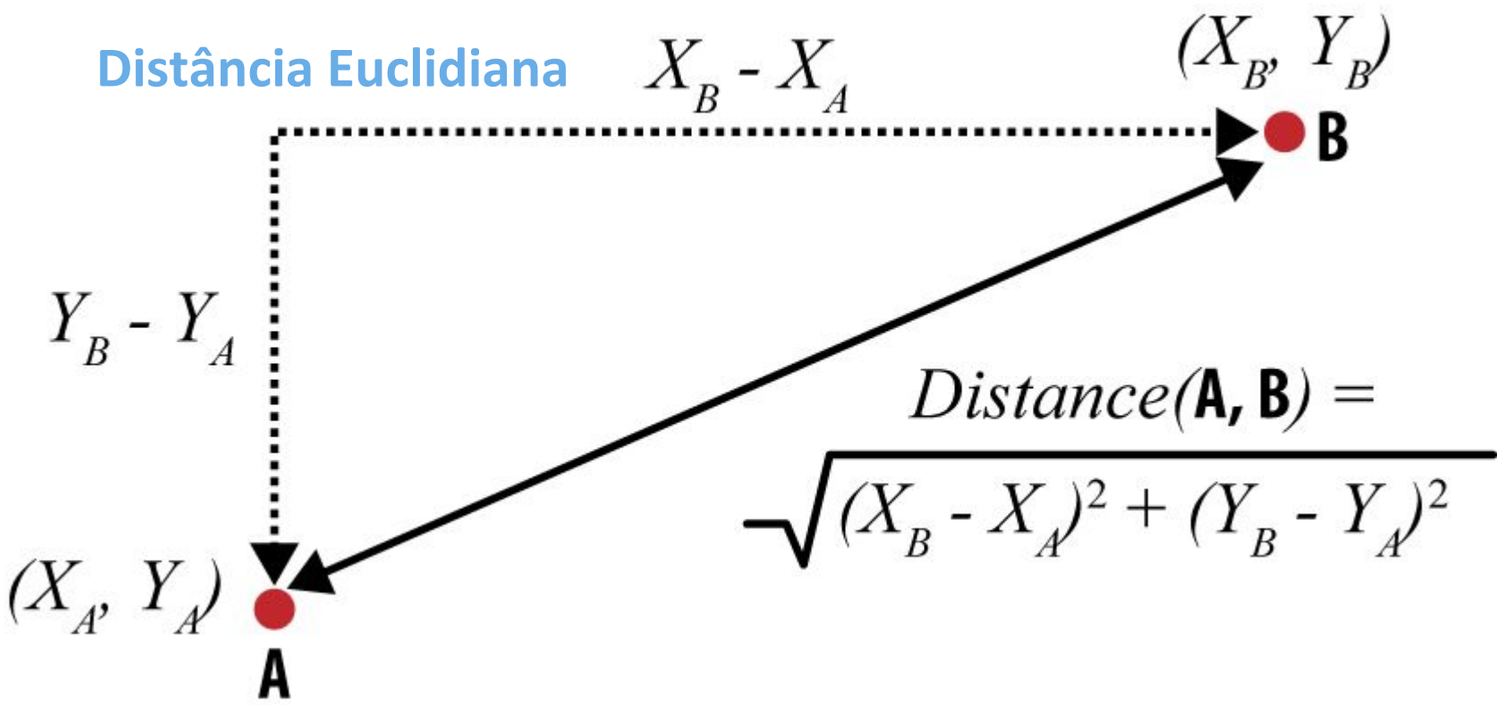
Atributo	Pessoa A	Pessoa B
Idade	23	40
Anos no endereço atual	2	10
Estado residencial (1=Proprietário, 2=Inquilino, 3=Outro)	2	1

Similaridade e Distância

- Se dois objetos são representados por vetores, é possível calcular a distância entre eles:

Atributo	Pessoa A	Pessoa B
Idade	23	40
Anos no endereço atual	2	10
Estado residencial (1=Proprietário, 2=Inquilino, 3=Outro)	2	1

Distância Euclidiana

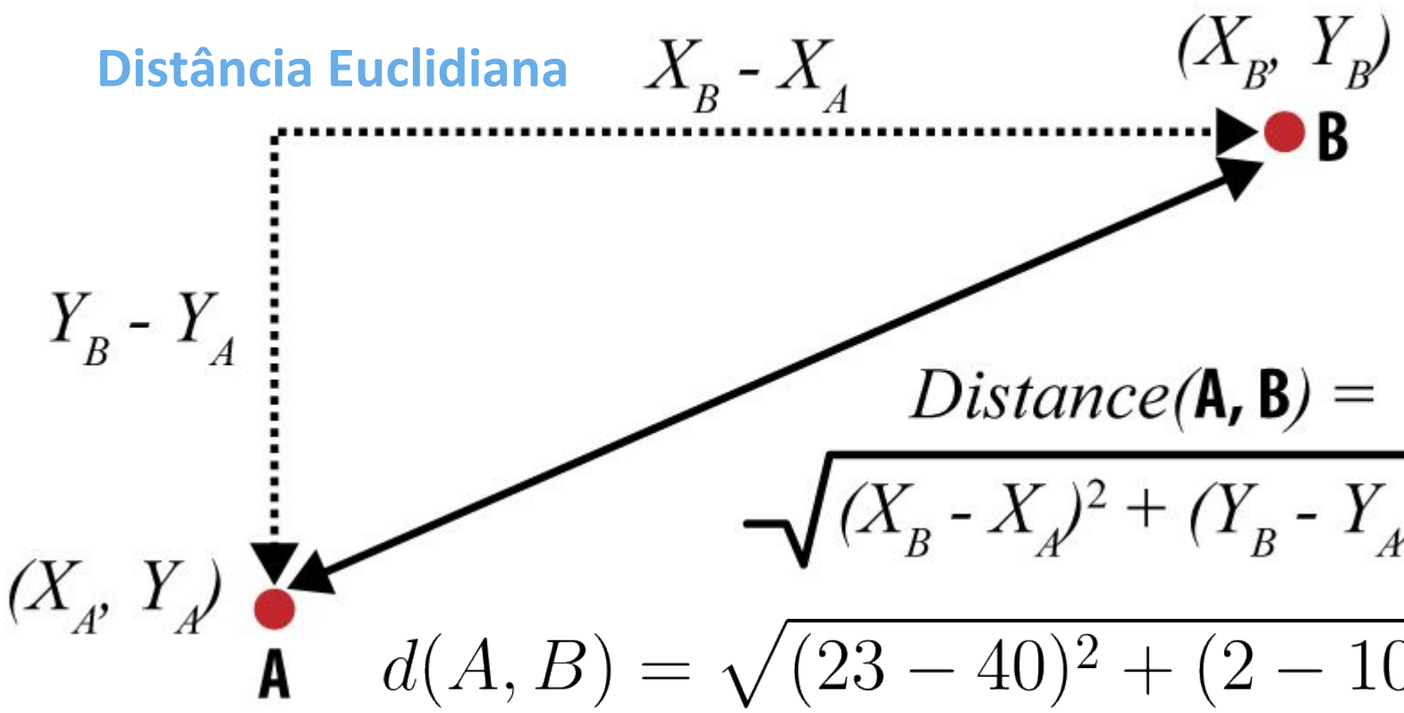


Similaridade e Distância

- Se dois objetos são representados por vetores, é possível calcular a distância entre eles:

Atributo	Pessoa A	Pessoa B
Idade	23	40
Anos no endereço atual	2	10
Estado residencial (1=Proprietário, 2=Inquilino, 3=Outro)	2	1

Distância Euclidiana



$$Distance(\mathbf{A}, \mathbf{B}) = \sqrt{(X_B - X_A)^2 + (Y_B - Y_A)^2}$$

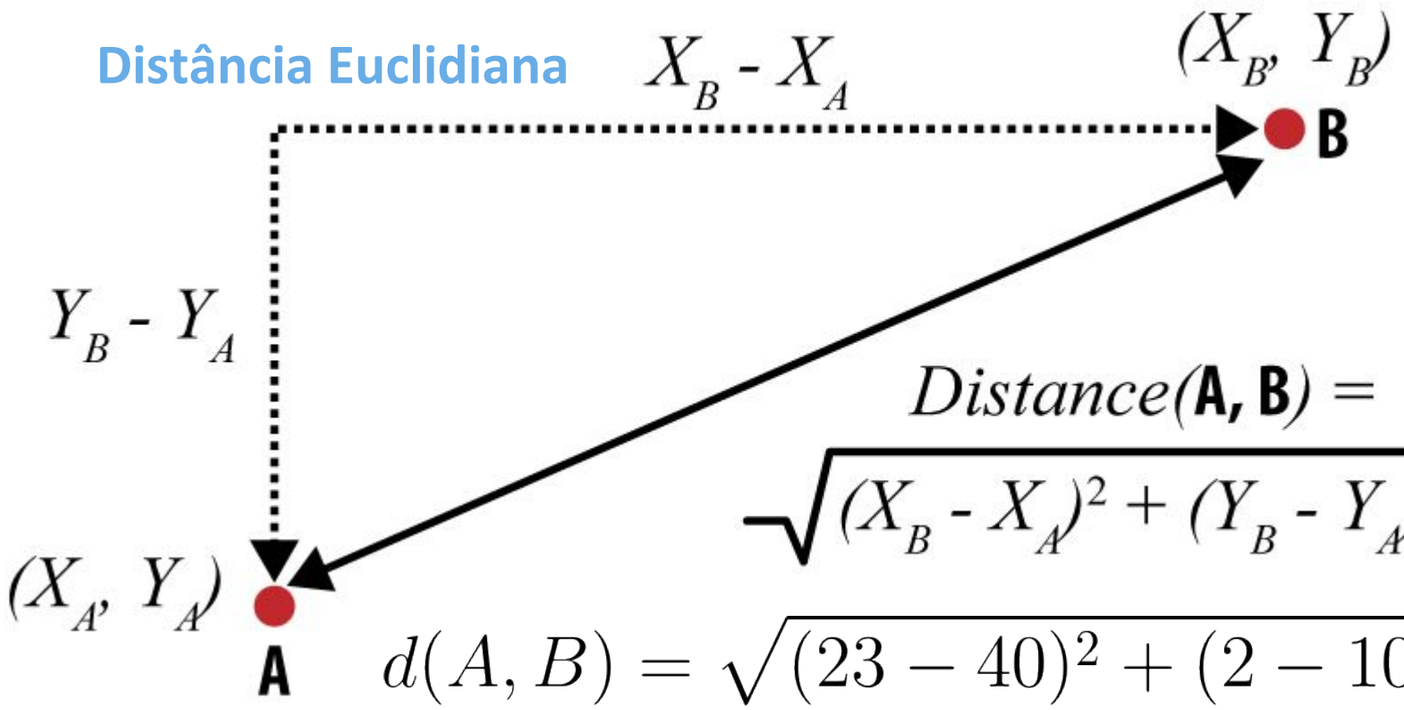
$$d(A, B) = \sqrt{(23 - 40)^2 + (2 - 10)^2 + (2 - 1)^2} = 18.8$$

Similaridade e Distância

- Se dois objetos são representados por vetores, é possível calcular a distância entre eles:

Atributo	Pessoa A	Pessoa B
Idade	23	40
Anos no endereço atual	2	10
Estado residencial (1=Proprietário, 2=Inquilino, 3=Outro)	2	1

Distância Euclidiana

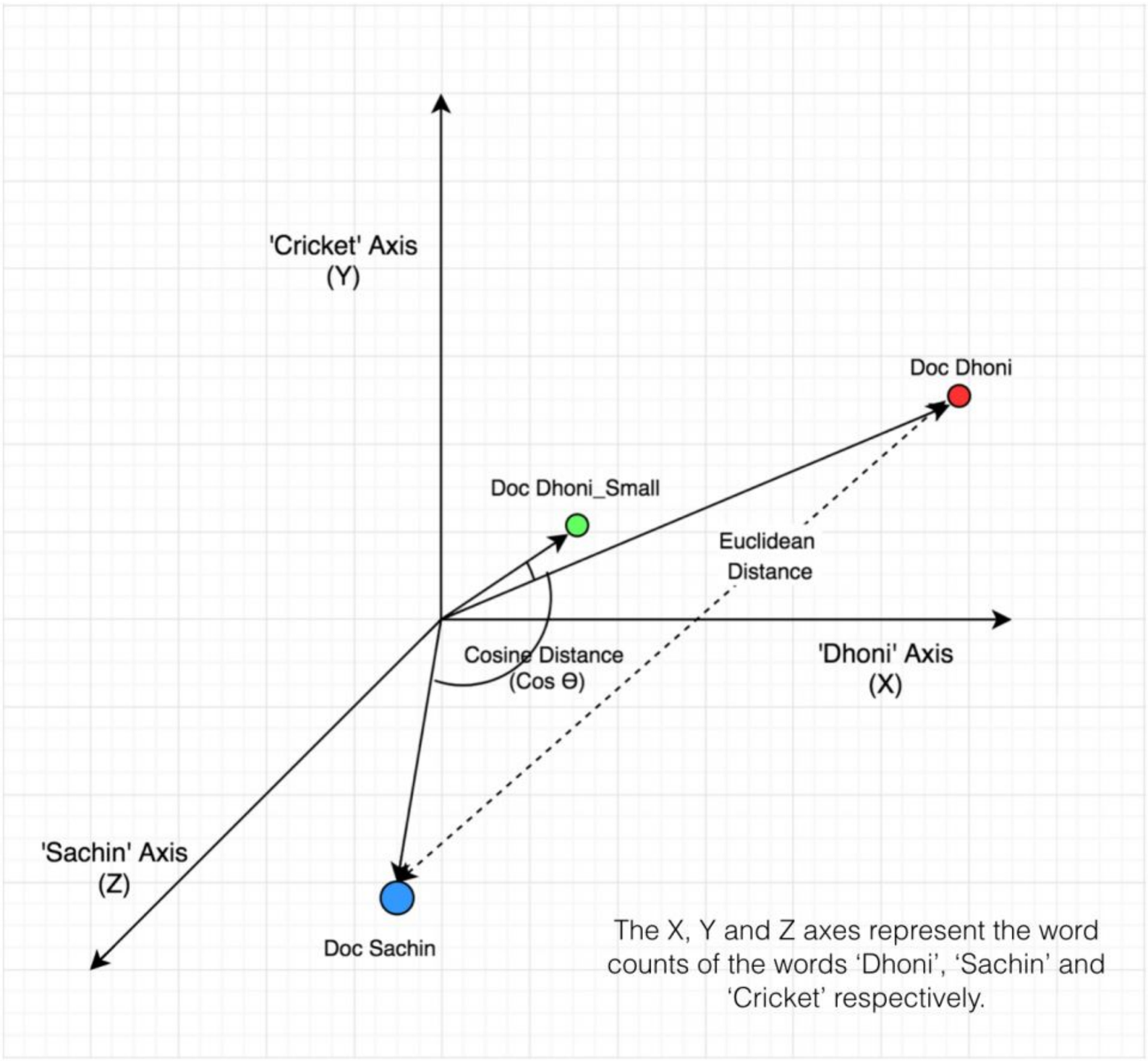


$$Distance(\mathbf{A}, \mathbf{B}) = \sqrt{(X_B - X_A)^2 + (Y_B - Y_A)^2}$$

$$d(A, B) = \sqrt{(23 - 40)^2 + (2 - 10)^2 + (2 - 1)^2} = 18.8$$

Normalmente a mais utilizada. É geral, intuitiva e computacionalmente muito rápida.

Outras funções de distância

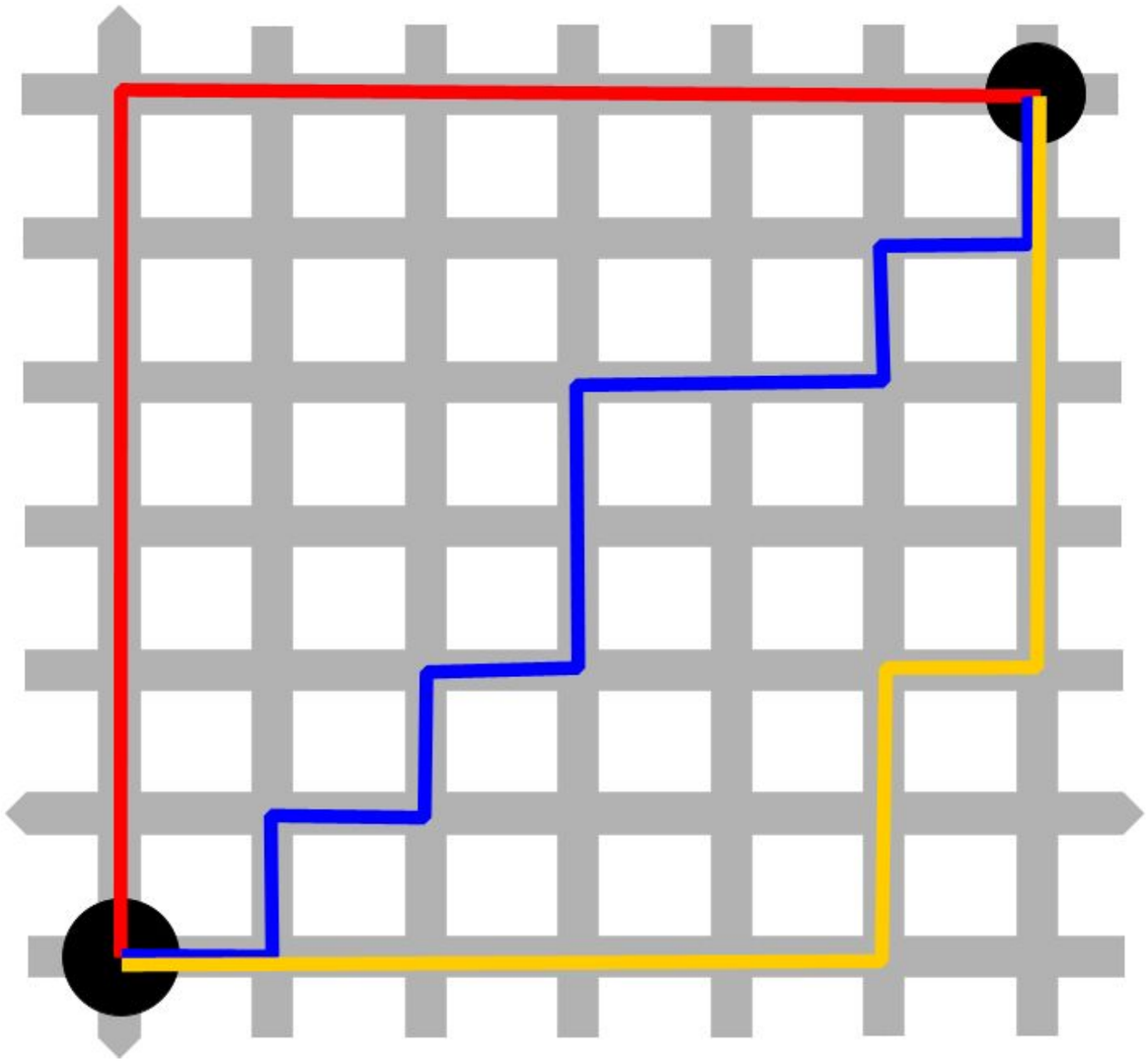


Distância de cosseno

$$d(A, B) = 1 - \frac{A \cdot B}{\|A\|_2 \cdot \|B\|_2}$$

Útil quando se quer ignorar diferenças de escala, como o tamanho dos textos

Outras funções de distância

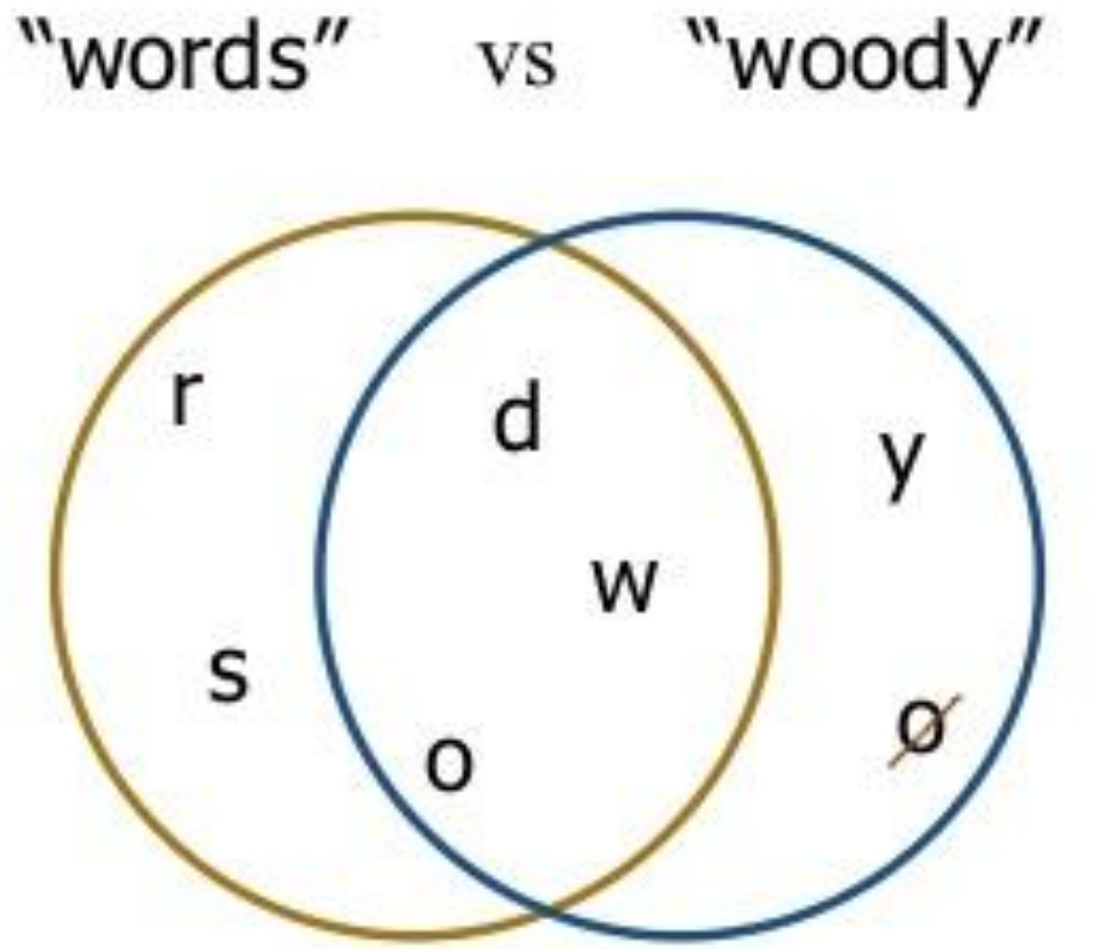


Distância de Manhattan

$$d(A, B) = \|A - B\|_1 = |x_1 - y_1| + |x_2 - y_2| + \ldots$$

Outras funções de distância

Distância de Jaccard



Trata dois objetos como **conjuntos** de características. Apropriado para problemas em que a posse de uma característica comum é importante, mas a ausência não.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{3}{7} = 0.43$$

6 0.5