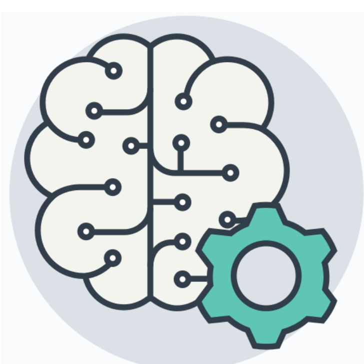
Aprendizado de Máquina

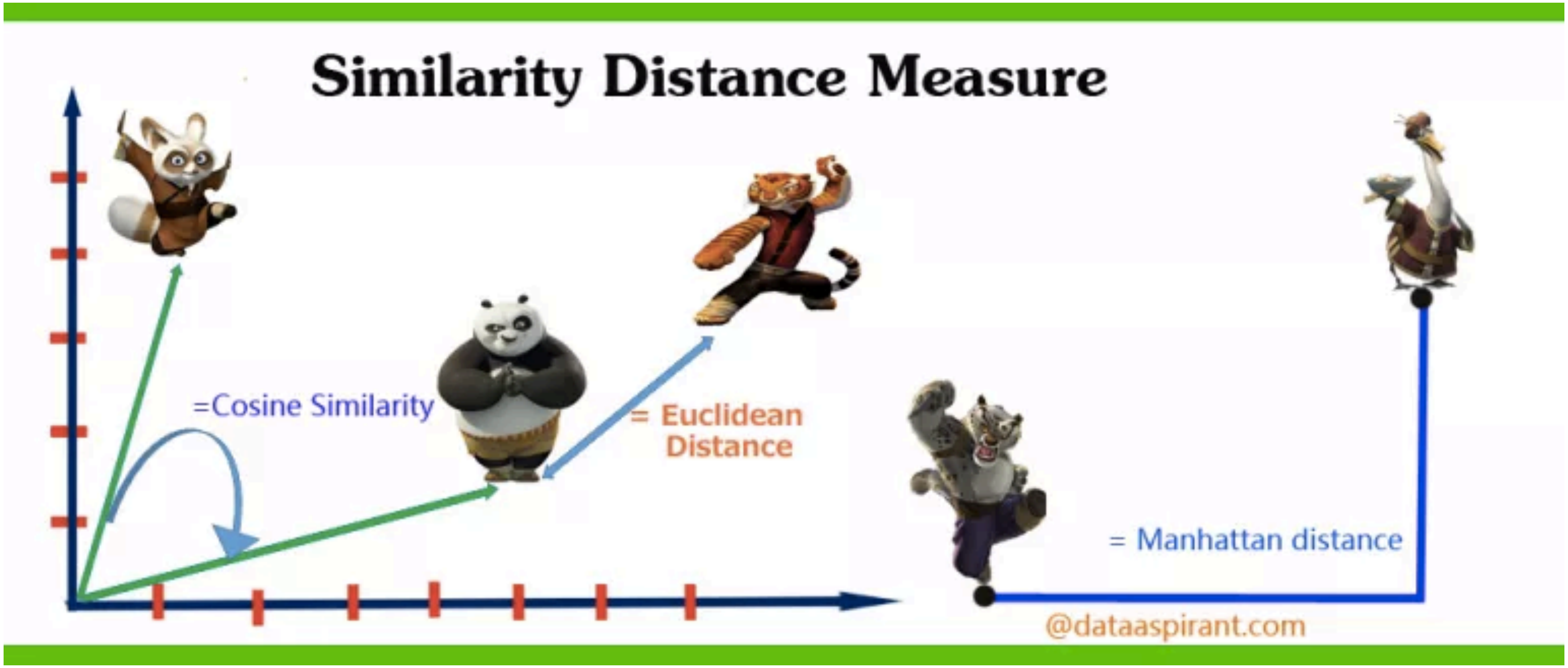
Medidas de Distância



Prof. Regis Pires Magalhães

regismagalhaes@ufc.br - http://bit.ly/ufcregis

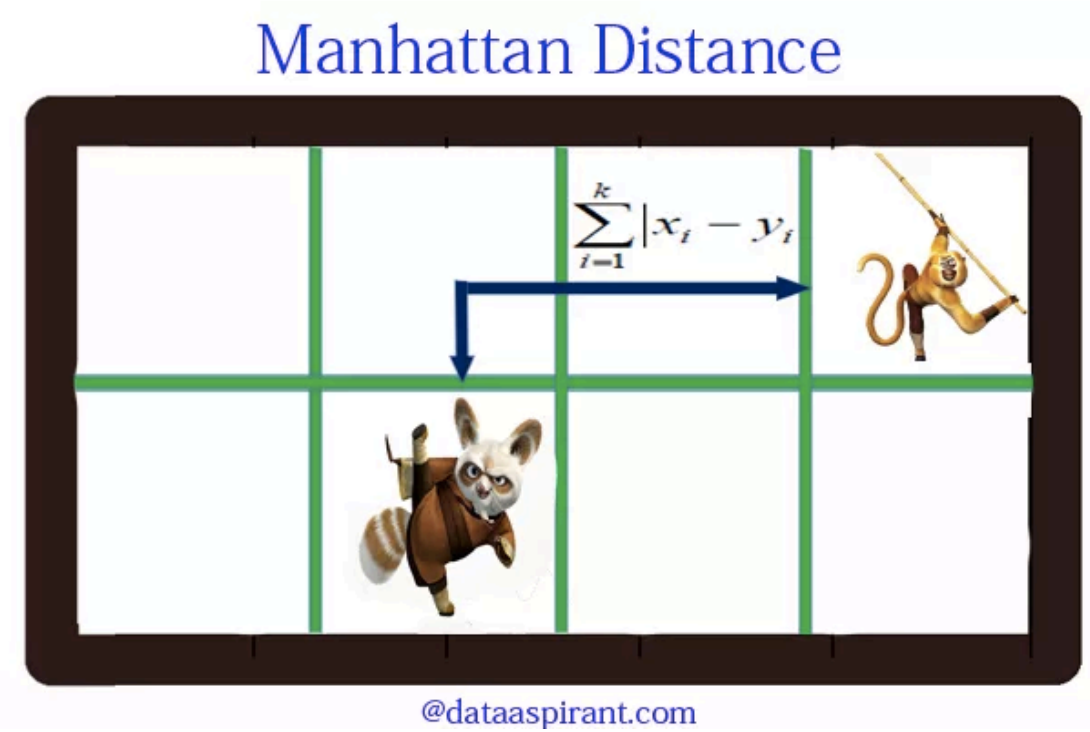
Medidas de similaridade / distância



Similaridade

- Mede quão parecidos 2 objetos são.
- Similarity are measured in the range 0 to 1 [0,1].

- Similarity = 1 if $X = Y$
- Similarity = 0 if $X \neq Y$
  - where X, Y are two objects.



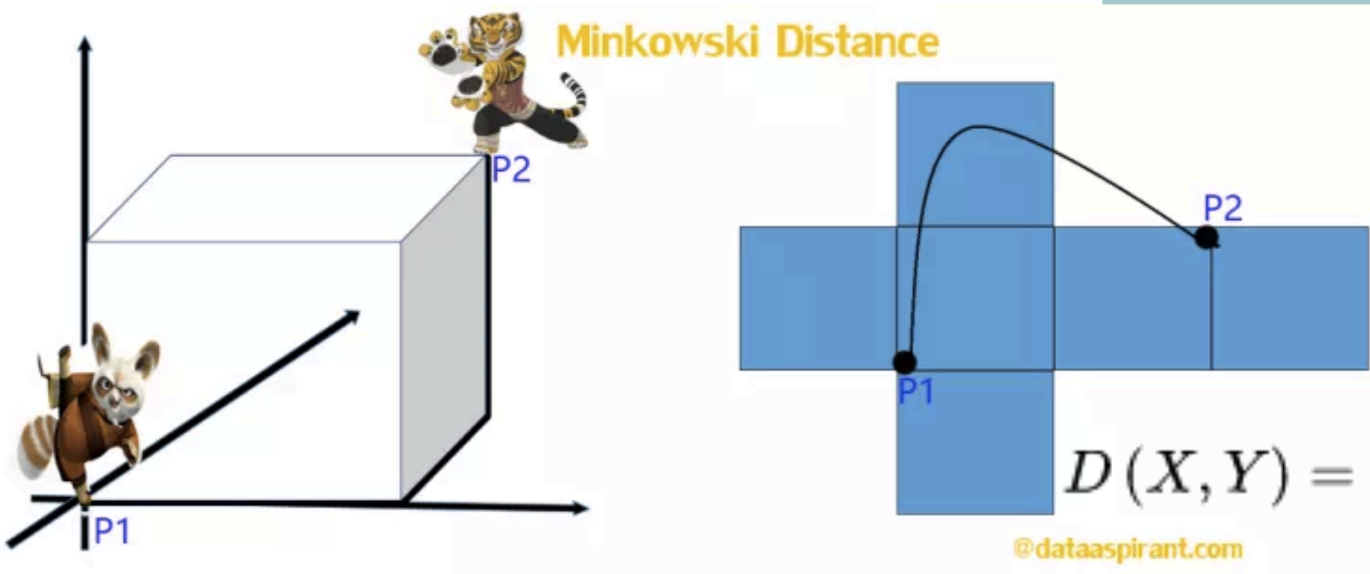
Sum of the absolute differences of their Cartesian coordinates.

Total sum of the difference between the x-coordinates and y-coordinates.

In a plane with p_1 at (x_1, y_1) and p_2 at (x_2, y_2).

Manhattan distance $= |x_1 - x_2| + |y_1 - y_2|$

Also known as Manhattan length, rectilinear distance, L1 distance or L1 norm, city block distance, Minkowski's L1 distance, taxi-cab metric, or city block distance.



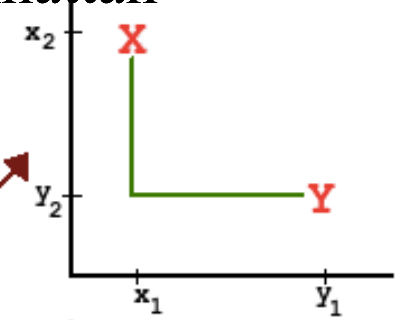
$$D(X,Y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{1/p}$$

Generalized metric form of Euclidean distance and Manhattan distance.

$$\text{Minkowski}(X,Y,\lambda) = \sqrt[\lambda]{\sum_{i=1}^{n} \left(w_i^{\lambda} * (|x_i - y_i|)^{\lambda} \right)}$$

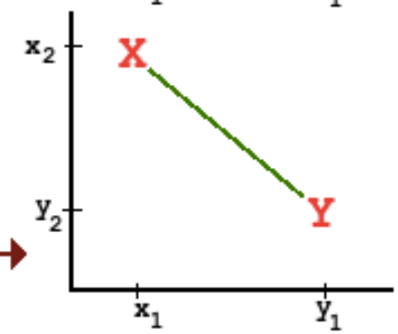
$\lambda = 1$

$$\text{City Block}(X,Y) = \sum_{i=1}^{n} \left(w_i * (|x_i - y_i|) \right)$$



$\lambda = 2$

$$\text{Euclidean}(X,Y) = \sqrt{\sum_{i=1}^{n} \left(w_i^{2} * (|x_i - y_i|)^{2} \right)}$$



Distance functions

Euclidean

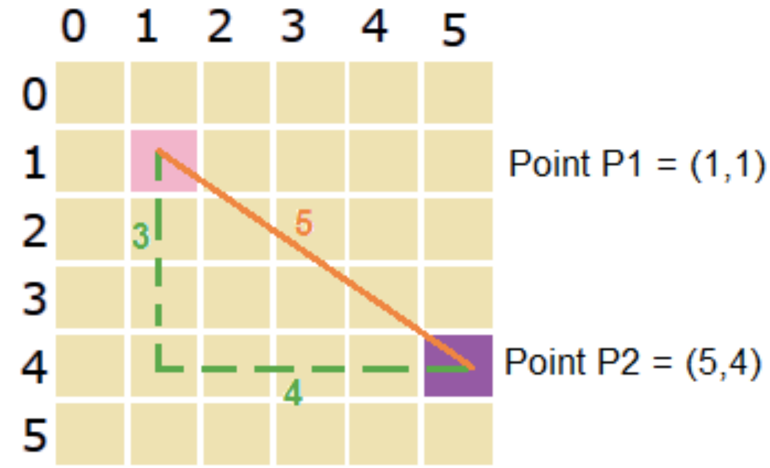
$$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$$

Manhattan

$$\sum_{i=1}^{k}|x_i - y_i|$$

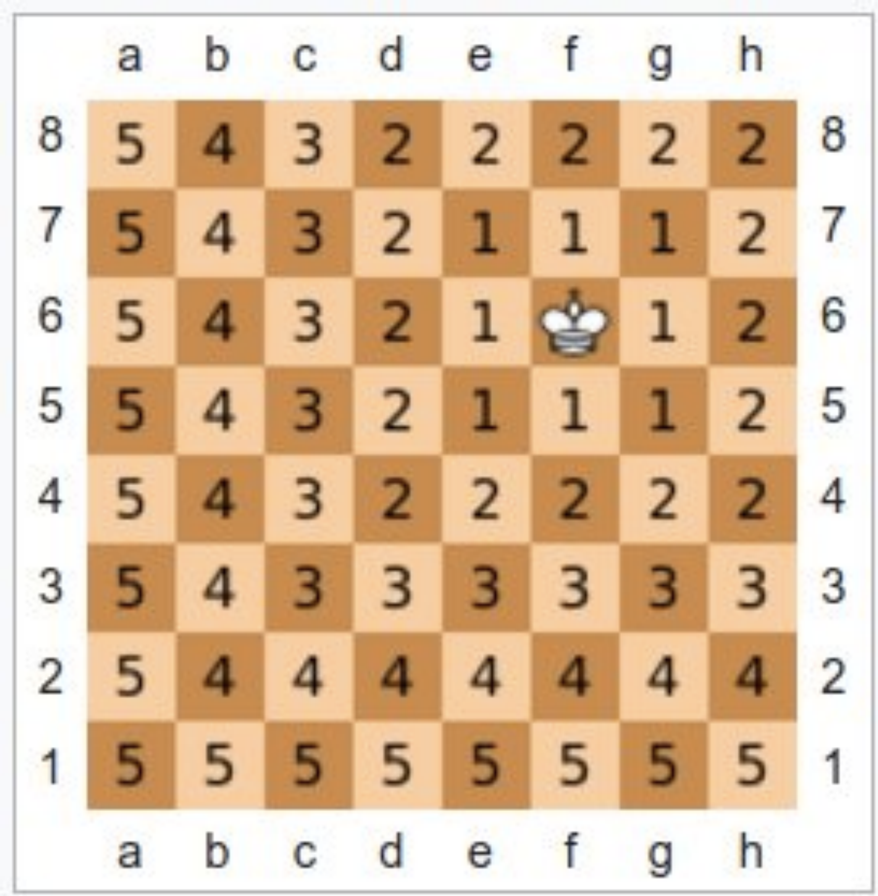
Minkowski

$$\left(\sum_{i=1}^{k}(|x_i - y_i|)^q\right)^{1/q}$$



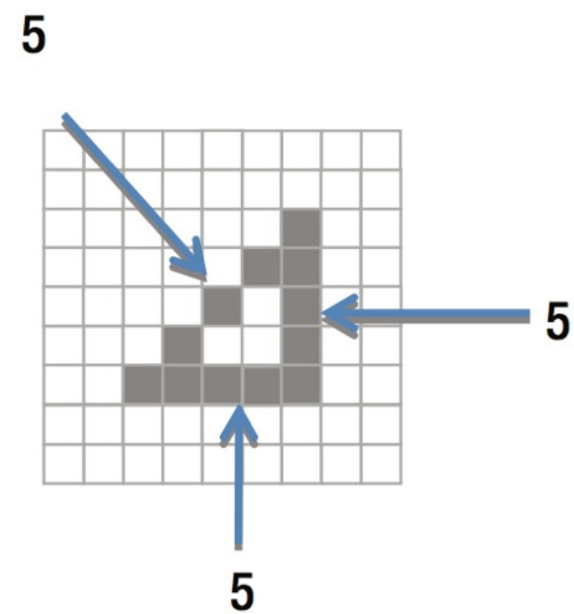
Euclidean distance = $\sqrt{(5-1)^2 + (4-1)^2} = 5$

Manhattan distance = $|5-1| + |4-1| = 7$

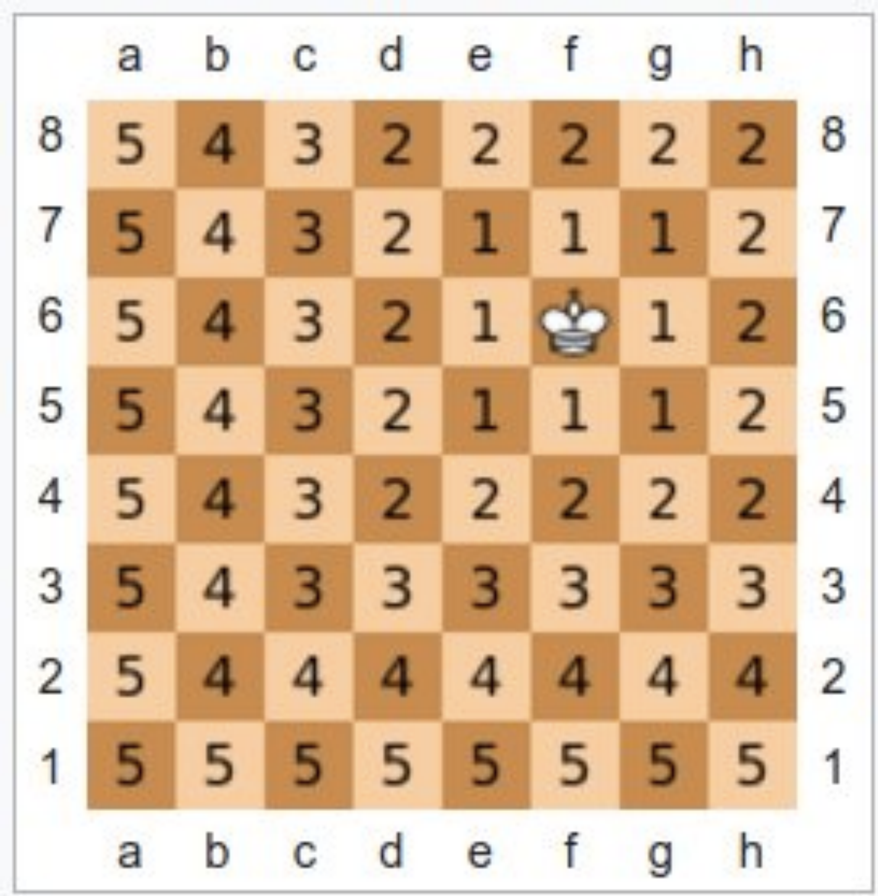


Chebyshev distance

also chessboard is best defined as a distance metric "where the distance between two vectors is the greatest of their differences along any coordinate dimension."



$$ChebyshevDistance\left[\{a,b\},\{x,y\}\right] = Max[Abs(a-x), Abs(b-y)]$$



No jogo de Xadrez:

Torres → Distância Manhatan

Reis e Rainhas → Distância Chebyshev

Bispos → Distância Manhatan (tabuleiro rotacionado 45º)

Canberra distance is a weighted version of Manhattan distance, which "has been used as a metric for comparing ranked lists and for intrusion detection in computer security."

$$d(\mathbf{p}, \mathbf{q}) = \sum_{i=1}^{n} \frac{|p_i - q_i|}{|p_i| + |q_i|}$$

where p and q are vectors and

$\mathbf{p} = (p_1, p_2, \dots, p_n)$ and $\mathbf{q} = (q_1, q_2, \dots, q_n)$

Hamming Distance

$$D_H = \sum_{i=1}^{k} |x_i - y_i|$$

$$x = y \Rightarrow D = 0$$

$$x \neq y \Rightarrow D = 1$$

X	Y	Distance
Male	Male	0
Male	Female	1

In the case of **categorical variables** you must use the **Hamming distance**.

Number of symbol changes necessarily to transform one to the other.

Inverted index (índice analítico)

doc1:"I like football"

doc2:"John likes football"

doc3:"John likes basketball"

I →	doc1	
like →	doc1	
football →	doc1	doc2
John →	doc2	doc3
likes →	doc2	doc3
football →	doc1	doc2
basketball →	doc3	

Consulta

- Who likes basketball?

  - Who:1 likes:1 basketball:1

I	→	doc1	
like	→	doc1	
football	→	doc1	doc2
John	→	doc2	doc3
likes	→	doc2	doc3
football	→	doc1	doc2
basketball	→	doc3	

  - Result: doc2, doc3 (without any order... or no?)

Boolean model

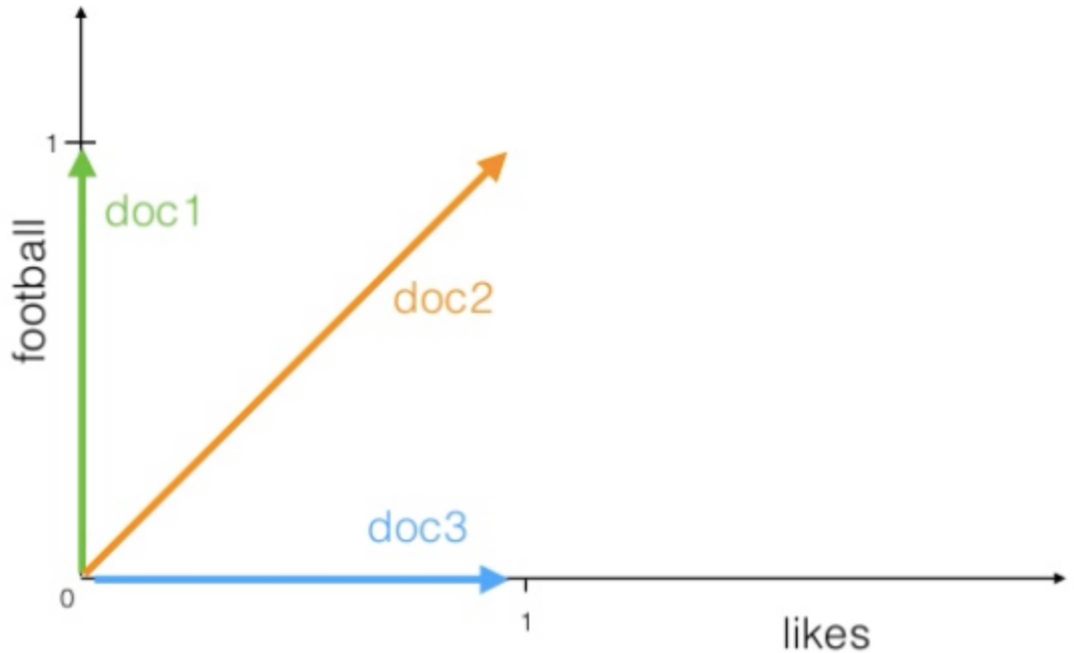
likes	→	doc2	doc3
basketball	→	doc3	

- Boolean Model: represents only presence or absence of query words and ranks document according to how many query words they contains

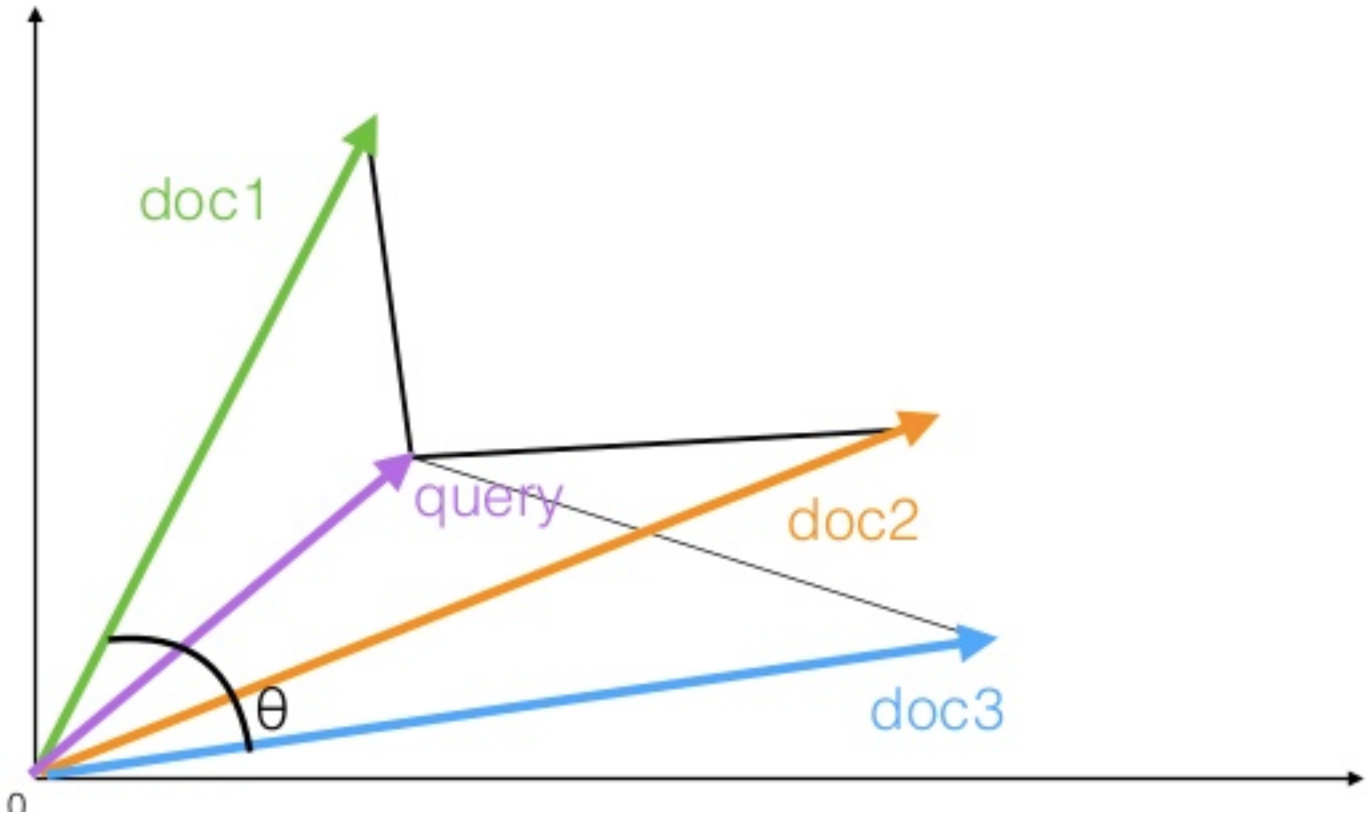
- Result: doc3, doc2 (with this prcise order)

Vector Space Model

	doc1	doc2	doc3
I	1	0	0
like	1	0	0
football	1	1	0
John	0	1	1
likes	0	1	1
football	1	1	0
basketball	0	0	1



Similarity between docs and queries



Cosine Distance / Similarity

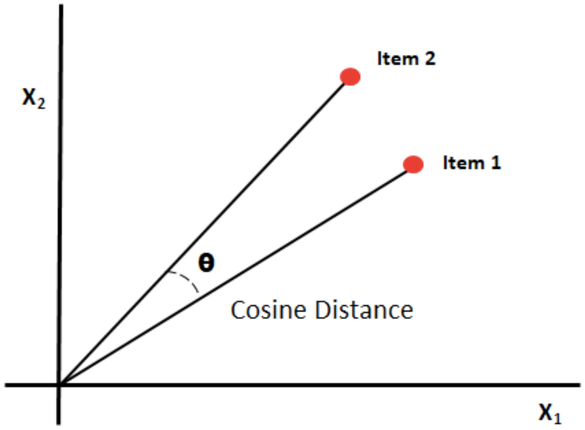
Two vectors with the same orientation have a cosine similarity of 1, two vectors at 90° have a similarity of 0, and two vectors diametrically opposed have a similarity of -1, independent of their magnitude.

$$\cos(\theta) = \frac{A.B}{\|A\|\,\|B\|}$$

The cosine of 0° is 1, and it is less than 1 for any other angle.

$$CosineDistance\left[\{a,b\},\{x,y\}\right] = 1 - \frac{ax+by}{\sqrt{a^2+b^2}\sqrt{x^2+y^2}}$$

Cosine Distance/Similarity

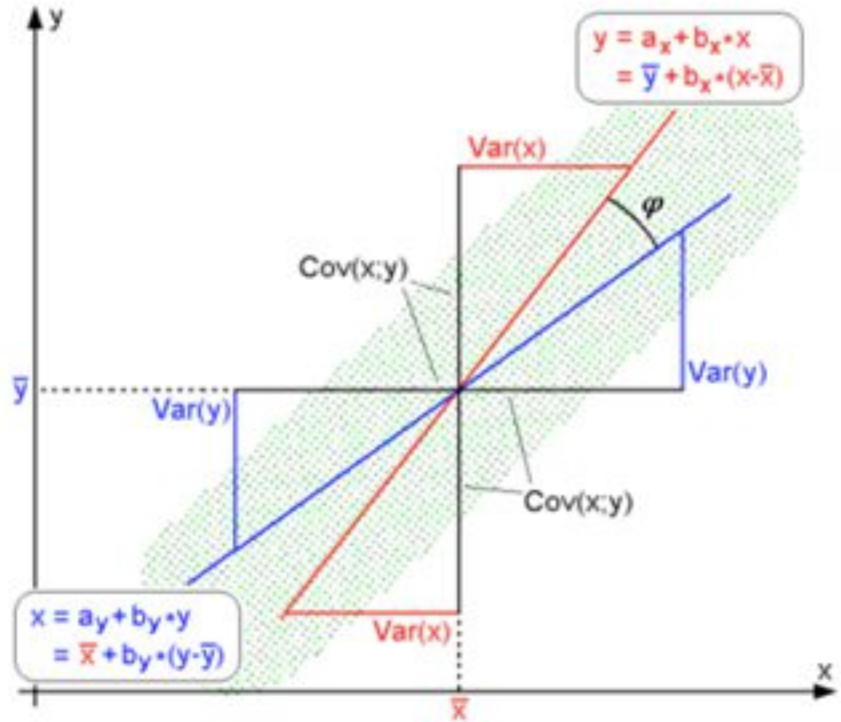


Cosine = Pearson

Pearson correlation is a measure of the correlation (linear dependence) between two variables X and Y, giving a value between +1 and −1 inclusive

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}.$$

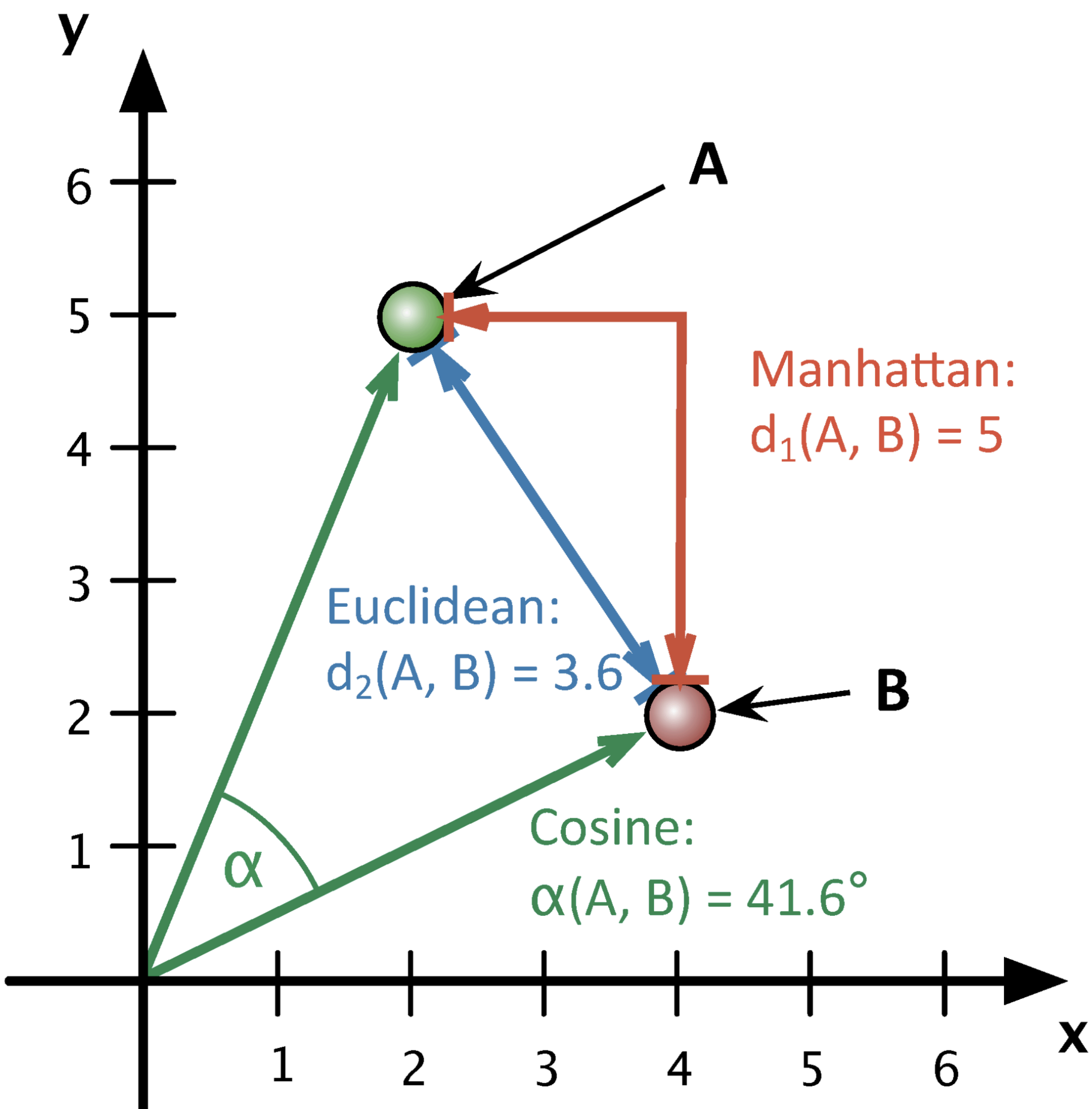
For uncentered data, the Pearson correlation coefficient corresponds with the the cosine of the angle ϕ between both possible regression lines $y=g_x(x)$ and $x=g_y(y)$.



Cosine similarity

- As the vectors represents documents and queries, the cosine is a measure of similarity of how similar is a document with respect to the query

- The documents obtained from the inverted index are ranked accordin to their cosine similarity wrt the query



y
6
5
4
3
2
1
A
B
Manhattan:
$d_1(A, B) = 5$
Euclidean:
$d_2(A, B) = 3.6$
Cosine:
$\alpha(A, B) = 41.6°$
α
1
2
3
4
5
6
x

Jaccard Similarity

Set A = {  }

Set B = {  }

|A| = 4 |B| = 5

Union(A,B) = {  }

Intersection (A,B) = {  }

|Union (A,B)| = 7 | Intersection (A,B)| = 2

Jaccard Similarity J (A,B) = | Intersection (A,B) | / | Union (A,B) |

= 2 / 7

= 0.286

Obrigado!

Dúvidas, comentários, sugestões?

Regis Pires Magalhães
regismagalhaes@ufc.br

VIRTVS VNITA FORTIOR
UFC