

# ***Big Data Analytics***

## Lecture 1 EXTRA – Cosine Similarity



# Cosine similarity



Let's assume the following DfM:

<i>Document</i>	<i>team</i>	<i>coach</i>	<i>hockey</i>	<i>baseball</i>	<i>soccer</i>	<i>penalty</i>	<i>score</i>	<i>win</i>	<i>loss</i>	<i>season</i>
<i>Document1</i>	5	0	3	0	2	0	0	2	0	0
<i>Document2</i>	3	0	2	0	1	1	0	1	0	1

As already highlighted, DfMs are typically very long and sparse (i.e., they have many 0 values). As a result, two term-frequency vectors may have many 0 values in common, meaning that the corresponding documents do not share many words, but this does not make them similar

We need a measure that will focus on the words that the two documents do have in common, and the occurrence frequency of such words. In other words, we need a measure for numeric data that ignores zero-matches.

# Cosine similarity



**Cosine similarity** is a measure of similarity that can be used to compare documents solving the problem just highlighted

Using the cosine measure as a similarity function, we have:

$$\text{similarity}(x, y) = \cos \theta = \frac{x \cdot y}{\sqrt{x_1^2 + x_2^2 + \dots + x_n^2} \cdot \sqrt{y_1^2 + y_2^2 + \dots + y_n^2}}$$

In our case:

Document1= $x=(5,0,3,0,2,0,0,2,0,0)$

Document2= $y=(3,0,2,0,1,1,0,1,0,1)$

$x \cdot y = (5 \cdot 3 + 0 \cdot 0 + 3 \cdot 2 + 0 \cdot 0 + 2 \cdot 1 + 0 \cdot 1 + 0 \cdot 0 + 2 \cdot 1 + 0 \cdot 0 + 0 \cdot 1) = 20$

$\sqrt{x_1^2 + x_2^2 + \dots + x_n^2} = 6.48$ ;  $\sqrt{y_1^2 + y_2^2 + \dots + y_n^2} = 4.12$

$\text{similarity}(x, y) = 0.94$

Conclusion: quite similar texts!