# *Big Data Analytics*

## Lab 1 EXTRA B – Cosine Similarity

UNIVERSITÄT LUZERN

# Cosine similarity

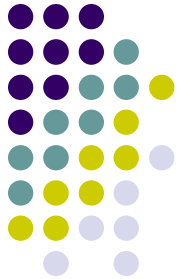
Let's assume the following DfM:

| Document | team | coach | hockey | baseball | soccer | penalty | score | win | loss | season |
|---|---|---|---|---|---|---|---|---|---|---|
| Document1 | 5 | 0 | 3 | 0 | 2 | 0 | 0 | 2 | 0 | 0 |
| Document2 | 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

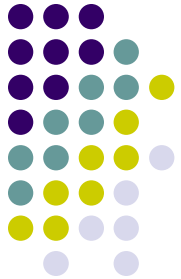


**Cosine similarity** is a measure of *semantic similarity* that can be used to compare documents solving the problem just highlighted

By representing texts as vectors in a Cartesian space via the DfM, cosine similarity estimates the differences between two texts based on *vectors of word occurrences*

More in details, the measure computes the *cosine of the angle between each vector x and y* (where x and y are two different texts)

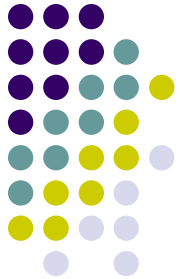
# Cosine similarity



A cosine value of 0 means that the two vectors are at 90 degrees to each other (orthogonal) and have no match. The closer the cosine value to 1, the smaller the angle and the greater the match between vectors

As a result, the cosine similarity between two texts ranges between 0 and 1, where 0 is reached when two texts are completely different and 1 is reached when two texts have identical feature proportions

# Cosine similarity



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

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

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·y=(5*3+0*0+3*2+0*0+2*1+0*1+0*0+2*1+0*0+0*1)=25

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

$similarity(x, y)$ =0.94

Conclusion: quite similar texts!