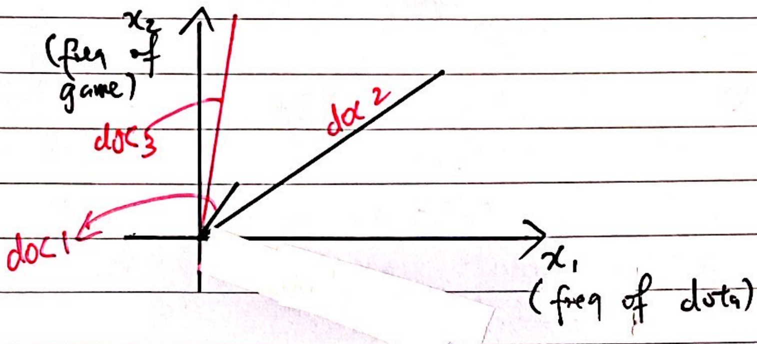
Cosine Similarity & Distance

Intuition:

2d-Dimension : We have 3 documents, and let us consider only 2 dominant terms in these 3 documents : dota, game. We represent the frequency of these 2 words as x_1 & x_2 in the 2-d plane.



doc 1 : (3, 6) = 3 appearances of dota & 6 counts of game
doc 2 : (12, 15) ⋮
doc 3 : (1, 10) ⋮

Now, Euclidean distance : $\|doc_1 - doc_2\|_2 = \sqrt{162}$, $\|doc_2 - doc_3\|_2 = \sqrt{146}$ } implies doc_2 is closer to doc_3 in Euclidean space.

However, in fact, both doc_1 & doc_2 are from the same extract : "Dota vs Humanity" while doc 3 is from "Games of the century". The catch is doc 3 is a super long article, and naturally (statisically) contain more words, → the chance of dota/game appearing is higher, but by right, doc 1 should be more similar.

Thus, in the 2d-space, in this situation, the euclidean magnitude may not be as informative. Instead, finding the angle between is better.

$$\cos\theta \text{ of } doc_1 \ \& \ doc_2 = \frac{\begin{bmatrix}3\\6\end{bmatrix}\begin{bmatrix}12\\15\end{bmatrix}}{\sqrt{45}\cdot\sqrt{369}} = 0.977$$

$$\cos\theta \text{ of } doc_2 \ \& \ doc_3 = \frac{\begin{bmatrix}12\\15\end{bmatrix}\begin{bmatrix}1\\10\end{bmatrix}}{\text{[illegible]}\cdot\sqrt{101}} = 0.839$$

DATE: NO:

Definition: Cosine Similarity

In vector space, given 2 vectors, the angle between them can be calculated as follows

$$\vec{A} \cdot \vec{B} = \|A\|_2 \|B\|_2 \cos\theta$$

iff

$$\theta = \cos^{-1} \frac{\vec{A} \cdot \vec{B}}{\|A\|_2 \|B\|_2}$$

iff

$$\cos\theta = \frac{\vec{A} \cdot \vec{B}}{\|A\|_2 \|B\|_2}$$

We then call $\cos\theta$ the "cosine similarity" of vector $\vec{A}$ & $\vec{B}$.

Note: $-1 \le \cos\theta \le 1$

-1 ↓ Exact opposite

1 ↓ Exactly the same

↓ if $\cos\theta = 0 \Rightarrow$ orthogonal.

$\boxed{\text{Cosine Distance}}$ = 1 − cosine similarity

$\boxed{\text{Cosine Similarity \& Euclidean Distance}}$

Given 2 vectors $\vec{A}, \vec{B}$, denote euclidean between A & B to be $\|A-B\|_2$, then

$$\|A-B\|_2^2 = (A-B)^T(A-B)$$

$$= \|A\|^2 + \|B\|^2 - 2A^TB$$

if A, B are normed.

$$= 2 - 2A^TB = 2(1 - \underbrace{A^TB}_{\text{cos similarity}})$$

$$= 2(1 - \cos(A, B))$$