Sentiment Analysis - VADER

https://ojs.aaai.org/index.php/ICWSM/article/view/14550/14399

Cosine Similarity

	Phone	good	not
S1	2	2	0
S2	1	1	1
S3	1	1	0

$\|S_1\| = \sqrt{2^2+2^2+0^2} = \sqrt{8}$

$\cos\theta_{S_3,S_2} = \frac{2}{\sqrt{3}\cdot\sqrt{2}}$

How do we measure similarity of two documents?

S1 = "Phone is good, phone is good"
S2 = "Phone is not good"
S3 = "It is a good phone"

$\sqrt{(2-1)^2+(2-1)^2+(0-0)^2} = \sqrt{2}$

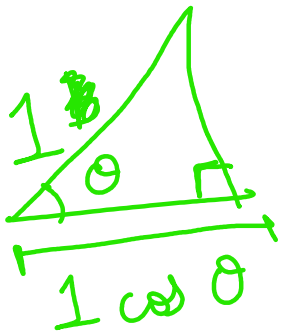
(Removing the stopwords except "not")
Euclidean distance from DTM would suggest that distance of S3 and S1 is $\sqrt{2}$, and distance of S3 and S2 is 1.

$\cos\theta_{S_1,S_3} = \frac{4}{\sqrt{8}\cdot\sqrt{2}} = \frac{4}{\sqrt{16}} = 1 \quad (1 < \sqrt{2})$

Cosine Similarity

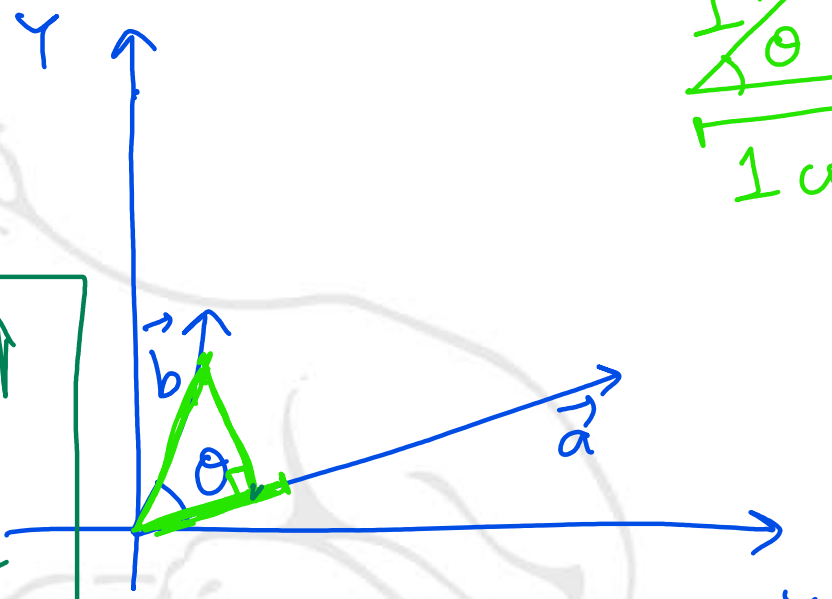
$$\cos\theta = \frac{\vec{a} \cdot \vec{b}}{\|\vec{a}\| \cdot \|\vec{b}\|}$$

$$\frac{\vec{a} \cdot \vec{b}}{\|\vec{a}\| \; \|\vec{b}\|}$$



$\theta \downarrow \rightarrow$ similar $\cos\theta \uparrow$

$\theta \uparrow \rightarrow$ not similar $\cos\theta \downarrow$



$$\|\vec{b}\| \cos\theta = \left(\frac{\vec{a}}{\|\vec{a}\|} \cdot \vec{b}\right)$$