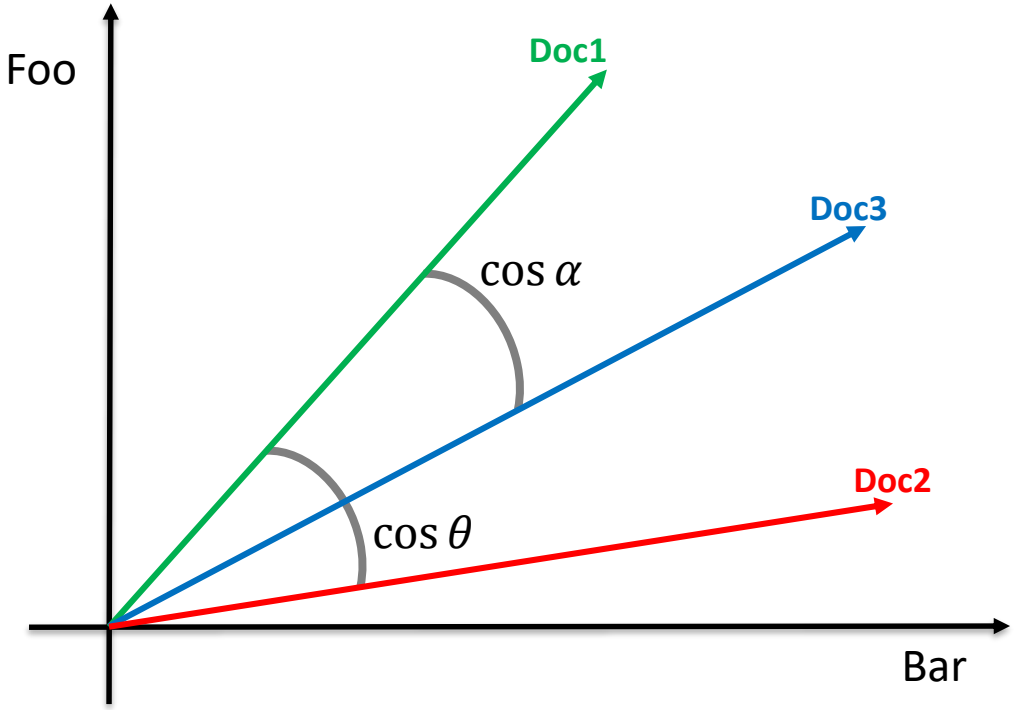
Similarity in Vector Space



Using the cosine between document vectors is an improvement over the dot product.

Advantages of using cosine for document similarity:

1. Given our representations, the cosine will be between [0, 1].
2. Metric works well in high dimensional spaces.

Cosine Similarity

Calculating cosine similarity:

$$\cos\theta = \frac{\mathrm{A} \cdot \mathrm{B}}{\|\mathrm{A}\|_2 \|\mathrm{B}\|_2} = \frac{\sum_{i=1}^{n} \mathrm{A}_i \mathrm{B}_i}{\sqrt{\sum_{i=1}^{n} \mathrm{A}_i^2} \sqrt{\sum_{i=1}^{n} \mathrm{B}_i^2}}$$

Given the document-term frequency matrix

bar	foo
6	10
10	3
8	7

Cosine similarity of Doc1 and Doc2

$$\frac{(6*10)+(10*3)}{\sqrt{(6^2)+(10^2)}\sqrt{(10^2)+(3^2)}} = \frac{60+30}{\sqrt{36+100}\sqrt{100+9}} = \frac{90}{\sqrt{136}\sqrt{109}} = 0.7391963$$

Cosine similarity of Doc1 and Doc3

$$\frac{(6*8)+(10*7)}{\sqrt{(6^2)+(10^2)}\sqrt{(8^2)+(7^2)}} = \frac{48+70}{\sqrt{36+100}\sqrt{64+49}} = \frac{118}{\sqrt{136}\sqrt{113}} = 0.9518606$$