Cosine Similarity

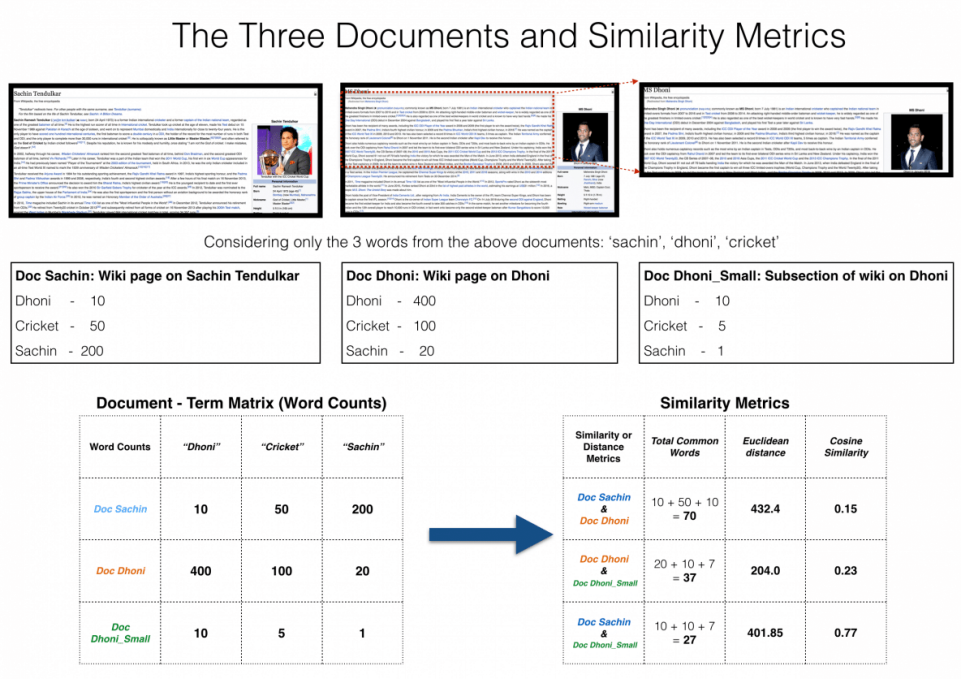
Friday, 21 February, 2020 10:38 AM

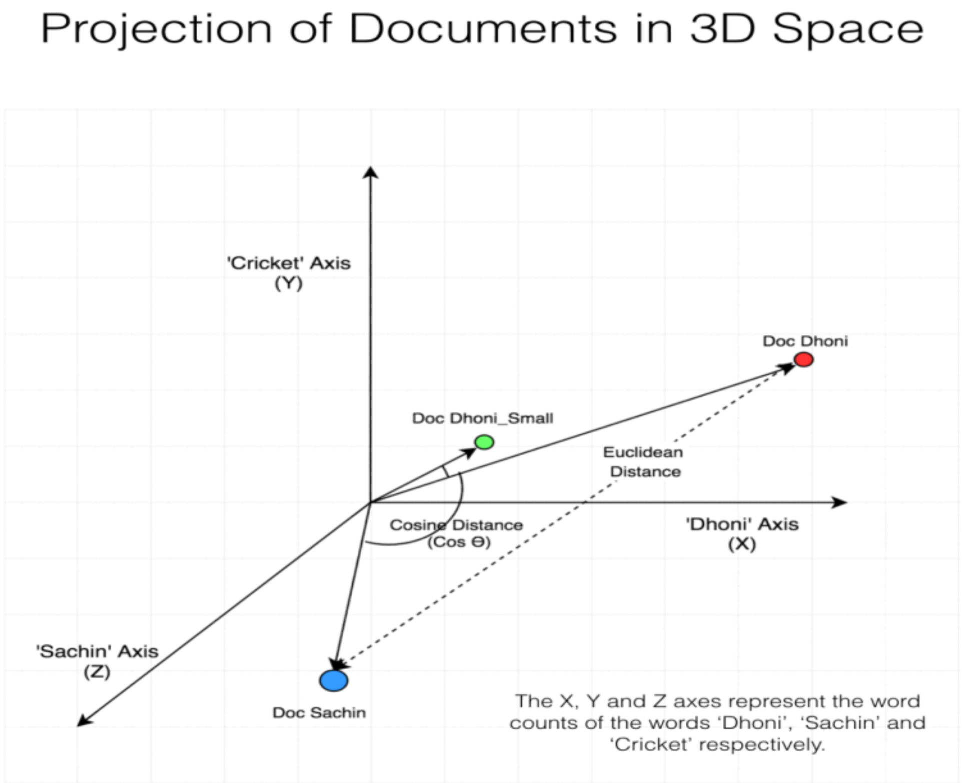
Cosine similarity is a metric used to measure how similar the documents are irrespective of their size. Mathematically, it measures the cosine of the angle between two vectors projected in a multi-dimensional space. The cosine similarity is advantageous because even if the two similar documents are far apart by the Euclidean distance (due to the size of the document), chances are they may still be oriented closer together. The smaller the angle, higher the cosine similarity.

From <https://www.machinelearningplus.com/nlp/cosine-similarity/>

Formula:

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}},$$




Projection of Documents in 3D Space
'Cricket' Axis
(Y)
Doc Dhoni
Doc Dhoni_Small
Euclidean
Distance
Cosine Distance
(Cos Θ)
'Dhoni' Axis
(X)
'Sachin' Axis
(Z)
Doc Sachin
The X, Y and Z axes represent the word counts of the words 'Dhoni', 'Sachin' and 'Cricket' respectively.

2`