Document distance Problem

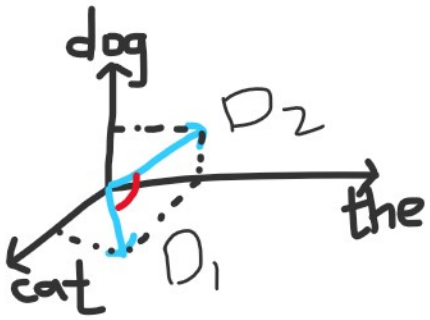
Wednesday, December 11, 2019 3:13 PM

document = sequence of words
word = string of alphanumeric chars.

Idea: shared words

D[w] = # occurrences of w in D

Example:



D1 = "the cat"
D2 = "the dog"

Create a vector...

Measure how different is vector D1 and D2...

How ?? ---> by dot product

But if there are million words in the document the dot product will be very huge.. Solution is to divide by the length of the vector... Take product of only common words ...

$$d(D1,D2) = \frac{D1 \cdot D2}{|D1| \cdot |D2|}$$

Algorithm:

1. Split document into words
2. Compute word frequencies
3. Compute dot product.