Fabian Gobet
MAI
19 Feb. 2024

Find the two closest documents based on their Freq feature vectors (use Euclidean distance), after having removed the stopwords {I, like, in, the, my, has, a, was}.

D1 = I like swimming in the sea
D2 = My car has a flat tire
D3 = A fish was swimming in the sea

$W' = \{$ I, like, swimming, in, the, sea, my, car, has, a, flat, tire, fish, was $\}$

$W = W' - \text{stopwords} = \{$ swimming, sea, car, flat, tire, fish $\}$

Frequency

$F_{D_1} = (1,1,0,0,0,0)$ $\quad F_{D_2} = (0,0,1,1,1,0)$

$F_{D_3} = (1,1,0,0,0,1)$

Normalized term-frequency

$TF_{D_1} = \left(\frac{1}{2}, \frac{1}{2}, 0, 0, 0, 0\right)$ $\quad TF_{D_2} = \left(0, 0, \frac{1}{3}, \frac{1}{3}, \frac{1}{3}, 0\right)$

$TF_{D_3} = \left(\frac{1}{3}, \frac{1}{3}, 0, 0, 0, \frac{1}{3}\right)$

Term frequency inverse-document-frequency

$TFIDF_{D_1} = \left(\frac{1}{2}\ln\left(\frac{3}{2}\right), \frac{1}{2}\ln\left(\frac{3}{2}\right), 0, 0, 0, 0\right)$

$TFIDF_{D_2} = \left(0, 0, \frac{1}{3}\ln\left(\frac{3}{1}\right), \frac{1}{3}\ln\left(\frac{3}{1}\right), \frac{1}{3}\ln\left(\frac{3}{1}\right), 0\right)$

$TFIDF_{D_3} = \left(\frac{1}{3}\ln\left(\frac{3}{2}\right), \frac{1}{3}\ln\left(\frac{3}{2}\right), 0, 0, 0, \frac{1}{3}\ln\left(\frac{3}{1}\right)\right)$