An Example of Vector Space Model

Dell Zhang
30/11/2006

Query

q: "gold silver truck"

Document Collection

d_1: "Shipment of gold arrived in a truck."
d_2: "Shipment of gold damaged in a fire."
d_3: "Delivery of silver arrived in a silver truck."

Term IDF Weights

The number of documents in the collection $n = 3$.

$idf_{a} = \log(n / df_{a}) = \log(3 / 3) = 0$
$idf_{arrived} = \log(n / df_{arrived}) = \log(3 / 2) = 0.18$
$idf_{damaged} = \log(n / df_{damaged}) = \log(3 / 1) = 0.48$
$idf_{delivery} = \log(n / df_{delivery}) = \log(3 / 1) = 0.48$
$idf_{fire} = \log(n / df_{fire}) = \log(3 / 1) = 0.48$
$idf_{gold} = \log(n / df_{gold}) = \log(3 / 2) = 0.18$
$idf_{in} = \log(n / df_{in}) = \log(3 / 3) = 0$
$idf_{of} = \log(n / df_{of}) = \log(3 / 3) = 0$
$idf_{shipment} = \log(n / df_{shipment}) = \log(3 / 2) = 0.18$
$idf_{silver} = \log(n / df_{silver}) = \log(3 / 1) = 0.48$
$idf_{truck} = \log(n / df_{truck}) = \log(3 / 2) = 0.18$

TF×IDF Document Vectors

$$w_{i,j} = tf_{i,j} \times idf_{i}$$

	a	arrived	damaged	delivery	fire	gold	in	of	shipment	silver	truck
d_1	0	0.18	0	0	0	0.18	0	0	0.18	0	0.18
d_2	0	0	0.48	0	0.48	0.18	0	0	0.18	0	0
d_3	0	0.18	0	0.48	0	0	0	0	0	0.96	0.18

Document Vector Length

$$\left|\vec{d}_{j}\right| = \sqrt{\sum_{i=1}^{m} w_{i,j}^{2}}$$

$$\left|\vec{d}_{1}\right| = \sqrt{0.18^{2} + 0.18^{2} + 0.18^{2} + 0.18^{2}} = 0.36$$

$$\left|\vec{d}_{2}\right| = \sqrt{0.48^{2} + 0.48^{2} + 0.18^{2} + 0.18^{2}} = 0.72$$

$$\left|\vec{d}_{3}\right| = \sqrt{0.18^{2} + 0.48^{2} + 0.96^{2} + 0.18^{2}} = 1.10$$

TF×IDF Query Vector

$$w_{i,j} = tf_{i,j} \times idf_{i}$$

	a	arrived	damaged	delivery	fire	gold	in	of	shipment	silver	truck
q	0	0	0	0	0	0.18	0	0	0	0.48	0.18

Query Vector Length

$$|\vec{q}| = \sqrt{\sum_{i=1}^{m} w_{i,q}^2}$$

$$|\vec{q}| = \sqrt{0.18^2 + 0.48^2 + 0.18^2} = 0.54$$

Query Processing with Cosine Similarities

$$sim(q, d_j) = \frac{\vec{q} \cdot \vec{d}_j}{|\vec{q}| \cdot |\vec{d}_j|} = \frac{\sum_{i=1}^{m} w_{i,q} w_{i,j}}{|\vec{q}| \cdot |\vec{d}_j|}$$

$$sim(q, d_1) = \frac{\sum_{i=1}^{11} w_{i,q} w_{i,1}}{|\vec{q}| \cdot |\vec{d}_1|}$$

$$= \frac{0\times0+0\times0.18+0\times0+0\times0+0\times0+0.18\times0.18+0\times0+0\times0+0\times0.18+0.48\times0+0.18\times0.18}{0.54\times0.36}$$

$$= \frac{0.18\times0.18+0.18\times0.18}{0.54\times0.36} = 0.33$$

$$sim(q, d_2) = \frac{\sum_{i=1}^{11} w_{i,q} w_{i,2}}{|\vec{q}| \cdot |\vec{d}_2|}$$

$$= \frac{0\times0+0\times0+0\times0.48+0\times0+0\times0.48+0.18\times0.18+0\times0+0\times0+0\times0.18+0.48\times0+0.18\times0}{0.54\times0.72}$$

$$= \frac{0.18\times0.18}{0.54\times0.72} = 0.08$$

$$sim(q, d_3) = \frac{\sum_{i=1}^{11} w_{i,q} w_{i,3}}{|\vec{q}| \cdot |\vec{d}_3|}$$

$$= \frac{0\times0+0\times0.18+0\times0+0\times0.48+0\times0+0.18\times0+0\times0+0\times0+0\times0+0.48\times0.96+0.18\times0.18}{0.54\times1.10}$$

$$= \frac{0.48\times0.96+0.18\times0.18}{0.54\times1.10} = 0.83$$

Search Result

Because $sim(q, d_3) > sim(q, d_1) > sim(q, d_2)$, the ranking of documents would be d_3, d_1, d_2.