Example of Similarity Coefficient Using Inner Product

Consider a case insensitive query and document collection with a query **Q** and a document collection consisting of the following three documents:

Q: "gold silver truck"
D_1: "Shipment of gold damaged in a fire"
D_2: "Delivery of silver arrived in a silver truck"
D_3: "Shipment of gold arrived in a truck"

In this collection, there are three documents, so $\boldsymbol{d} = \mathbf{3}$. If a term appears in only one of the three documents, its ***idf*** is $\log\frac{d}{df_j} = \log\frac{3}{1} = 0.477$. Similarly, if a term appears in two of the three documents its ***idf*** is $\log\frac{d}{df_j} = \log\frac{3}{2} = 0.176$, and a term which appears in all three documents has an ***idf*** of $\log\frac{3}{3} = 0$.

The ***idf*** for the terms in the three documents is given below:

$idf_a = 0$	$idf_{in} = 0$
$idf_{arrived} = 0.176$	$idf_{of} = 0$
$idf_{damaged} = 0.477$	$idf_{silver} = 0.477$
$idf_{delivery} = 0.477$	$idf_{shipment} = 0.176$
$idf_{fire} = 0.477$	$idf_{truck} = 0.176$
$idf_{gold} = 0.176$	

Document vectors can now be constructed. Since eleven terms appear in the document collection, an eleven-dimensional document vector is constructed. The alphabetical ordering given above is used to construct the document vector so that t_1 corresponds to term number one which is a and t_2 is arrived, etc. The weight for term i in vector j is computed as the $idf_i \times tf_{ij}$. The document vectors are shown in Table 2.1.

Table 2.1. Document Vectors

docid	a	arrived	damaged	delivery	fire	gold	in	of	shipment	silver	truck
D_1	0	0	.477	0	.477	.176	0	0	.176	0	0
D_2	0	.176	0	.477	0	0	0	0	0	.954	.176
D_3	0	.176	0	0	0	.176	0	0	.176	0	.176
Q	0	0	0	0	0	.176	0	0	0	.477	.176

$$\begin{aligned} SC(Q, D_1) &= (0)(0) + (0)(0) + (0)(0.477) + (0)(0) \\ &+ (0)(0.477) + (0.176)(0.176) + (0)(0) + (0)(0) \\ &+ (0)(0.176) + (0.477)(0) + (0.176)(0) \\ &= (0.176)^2 \approx 0.031 \end{aligned}$$

Similarly,

$$SC(Q, D_2) = (0.954)(0.477) + (0.176)^2 \approx 0.486$$

$$SC(Q, D_3) = (0.176)^2 + (0.176)^2 \approx 0.062$$

Hence, the ranking would be D_2, D_3, D_1.