

## An Example of Text Classification with $k$ NN ( $k=1$ )

Dell Zhang  
17/11/2006

Two Classes: ham and spam

### Training Data

ham  $d_1$ : "Shipment of gold arrived in a truck."

spam  $d_2$ : "Shipment of gold damaged in a fire."

### Test Data

?  $d_3$ : "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

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

$$|\vec{d}_1| = \sqrt{0.18^2 + 0.18^2 + 0.18^2 + 0.18^2} = 0.36$$

$$|\vec{d}_2| = \sqrt{0.48^2 + 0.48^2 + 0.18^2 + 0.18^2} = 0.72$$

$$|\vec{d}_3| = \sqrt{0.18^2 + 0.48^2 + 0.96^2 + 0.18^2} = 1.10$$

### // Document Cosine Similarities

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

$$\begin{aligned}
sim(d_3, d_1) &= \frac{\sum_{i=1}^{11} w_{i,3} w_{i,1}}{|\vec{d}_3| \cdot |\vec{d}_1|} \\
&= \frac{0 \times 0 + 0.18 \times 0.18 + 0 \times 0 + 0.48 \times 0 + 0 \times 0 + 0 \times 0.18 + 0 \times 0 + 0 \times 0 + 0 \times 0.18 + 0.96 \times 0 + 0.18 \times 0.18}{1.10 \times 0.36} \\
&= \frac{0.18 \times 0.18 + 0.18 \times 0.18}{1.10 \times 0.36} = 0.16 > 0
\end{aligned}$$

$$\begin{aligned}
sim(d_3, d_2) &= \frac{\sum_{i=1}^{11} w_{i,3} w_{i,2}}{|\vec{d}_3| \cdot |\vec{d}_2|} \\
&= \frac{0 \times 0 + 0.18 \times 0 + 0 \times 0.48 + 0.48 \times 0 + 0 \times 0.48 + 0 \times 0.18 + 0 \times 0 + 0 \times 0 + 0 \times 0.18 + 0.96 \times 0 + 0.18 \times 0}{1.10 \times 0.72} \\
&= \frac{0}{1.10 \times 0.72} = 0
\end{aligned}$$

// kNN Classification (k=1)

The nearest neighbour of  $d_3$  is  $d_1$  because  $sim(d_3, d_1) > sim(d_3, d_2)$ .

Since the class of  $d_1$  is **ham**,  $d_3$  should be classified into the **ham** class.