# Bag of Words

-Anurag Dhaipule.

(1) Document Similarity

let the representation be

[woof meow squeak]

Then

$$D1 = [2 \quad 1 \quad 0]$$
$$D2 = [2 \quad 0 \quad 1]$$

Using
$$Sim(A,B) = \frac{\sum_i a_i b_i}{\sqrt{\sum_i a_i^2}\sqrt{\sum_i b_i^2}} \qquad \text{— (1)}$$

a)
$$Sim(D1, D2) = \frac{2 \cdot 2 + 1 \cdot 0 + 0 \cdot 1}{\sqrt{2^2+1^2+0^2} \cdot \sqrt{2^2+0^2+1}}$$
$$= \frac{4}{5}.$$

b) Including idf weights
$$idf_i = \log\left(\frac{N}{n_i}\right)$$

$$idf_{\cancel{meow}\,woof} = \log \frac{2}{2} = 0$$
$$idf_{meow} = \log \frac{2}{1} = \log 2$$
$$idf_{squeak} = \log \frac{2}{1} = \log 2.$$

Now, the new representations for the documents will be

$D_1' = [2 \times idf_{woof}, \; 1 \times idf_{meow} \; \; 0 \times idf_{squeak}]$

$D_2' = [2 \times idf_{woof}, \; 0 \times idf_{meow} \; \; 1 \times idf_{squeak}]$

$D_1' = [0 \quad \log 2 \quad 0]$

$D_2' = [0 \quad 0 \quad \log 2]$

Using eq ①.

$$Sim(D_1', D_2') = \frac{0 \cdot 0 + 0 \cdot \log 2 + \log 2 \cdot 0}{\sqrt{0+(\log 2)^2+0} \; \sqrt{0+0+(\log 2)^2}}$$

$$= 0.$$

b) $D_3 = [0 \quad 1 \quad 1]$

Now $N = 3.$

New idfs, $idf_{woof} = \log \frac{3}{2}$

$idf_{meow} = \log \frac{3}{2}$

$idf_{squeek} = \log \frac{3}{2}$

New representations

$D_1' = [2 \times idf_{woof}, \; 1 \times idf_{meow}, \; 0 \times idf_{squeak}]$

$= [2 \log \frac{3}{2}, \; \log \frac{3}{2}, \; 0]$

similarly,

$D_2' = [2 \log \frac{3}{2}, \; 0, \; \log \frac{3}{2}]$

Using (1),

$$Sim(D_1', D_2') = \frac{2\log 3/2 \times 2\log 3/2 + 0 + 0}{\sqrt{(2\log\frac{3}{2})^2 + (\log 3/2)^2 + 0}\ \sqrt{(2\log 3/2)^2 + 0 + (\log 3/2)^2}}$$

$$= \frac{4(\log 3/2)^2}{5(\log 3/2)^2} = \frac{4}{5}$$

(2.) Naive Bayes and Smoothing.

(a) $$P(t) = \frac{\text{Count of docs labelled } t}{\text{Total docs}}$$

$P(+) = \frac{5}{10} = \frac{1}{2}$ $\qquad$ $P(-) = \frac{5}{10} = \frac{1}{2}$

$$P(w_i|t) = \frac{\text{Count of docs labelled } t \text{ containing } w_i}{\text{Count of docs labelled } t}$$

$P(great|+) = \frac{5}{5} = 1$ $\qquad$ $P(great|-) = \frac{0}{5} = 0$

$P(food|+) = \frac{5}{5} = 1$ $\qquad$ $P(food|-) = \frac{5}{5} = 1$

$P(served|+) = \frac{0}{5} = 0$ $\qquad$ $P(served|-) = \frac{1}{5} = \frac{1}{5}$

$P(terrible|+) = \frac{0}{5} = 0$ $\qquad$ $P(terrible|-) = \frac{5}{5} = 1$

Using Bernoulli's Naive Bayes

① $P(+ \mid \text{"great food served"}) = P(\text{great} \mid +) \times P(\text{food} \mid +) \times P(\text{served} \mid +) \times (1 - P(\text{terrible} \mid +)) \times P(+)$

$= 1 \times 1 \times 0 \times (1-0) \times \frac{1}{2} =$

$= 0.$

② $P(- \mid \text{"great food served"}) = P(\text{great} \mid -) * P(\text{food} \mid -) * P(\text{served} \mid -) * (1 - P(\text{terrible} \mid -)) * P(-)$

$= 0 * 1 * 1/5 * (1-1) * 1/2$

$= 0.$

(b) With laplace smoothing,
~~we~~ add

$$P(t) = \frac{\text{Count of doc labelled } t + 1}{\text{total doc} + 2}$$

$P(+) = \frac{6}{12} = \frac{1}{2}$ $\qquad P(-) = \frac{1}{2}$.

$|||^{ly}$

$P(\text{great} \mid +) = \frac{5+1}{5+2} = \frac{6}{7}$ ; $P(\text{great} \mid -) = \frac{1}{7}$

$P(\text{food} \mid +) = \frac{5+1}{5+2} = \frac{6}{7}$ ; $P(\text{food} \mid -) = \frac{6}{7}$

$P(\text{served} \mid +) = \frac{0+1}{5+2} = \frac{1}{7}$ ; $P(\text{served} \mid -) = \frac{2}{7}$

$P(\text{terrible} \mid +) = \frac{0+1}{5+2} = \frac{1}{7}$ $\qquad P(\text{served} \mid -) = \frac{6}{7}$

Using the equation ① & ②, we have.

$$P(+|\text{"great food served"}) = \frac{6}{7} \times \frac{6}{7} \times \frac{1}{7} \times \left(1 - \frac{1}{7}\right) \times \frac{1}{2}$$

$$= 0.045.$$

$$P(-|\text{"great food served"}) = \frac{1}{7} \times \frac{6}{7} \times \frac{2}{7} \times \left(1 - \frac{6}{7}\right) \times \frac{1}{2}$$

$$= 0.002$$

$P(+|\text{"great food served"})$ is greater.