CSCI.GA.2590 - Natural Language Processing – Spring 2016 – Prof. Grishman

Assignment #1

(1) [1 point] [document similarity] Suppose our pets have produced two documents
D1 = [woof woof meow]
D2 = [woof woof squeak]
(a) What is the cosine similarity of D1 and D2, not using idf weighting?
(b) What is the cosine similarity if idf weighting is used?
(c) How would the answer to (b) change if we added a third document
D3 = [meow squeak]
to the collection?

Solution:

(a) Feature vector $v = [woof, meow, squeak]$, then $tf_1 = [2, 1, 0]$ and $tf_2 = [2, 0, 1]$.
So $cos\_sim(D1, D2) = \frac{tf_1 \cdot tf_2}{|tf_1| \cdot |tf_2|} = 0.8$.

(b) $idf_{woof} = log\frac{N}{n_{dog}} = log\frac{2}{2} = 0$,
$idf_{meow} = log\frac{N}{n_{meow}} = log\frac{2}{1} = 0.3$,
$idf_{squeak} = log\frac{N}{n_{squeak}} = log\frac{2}{1} = 0.3$,
According to $w_i = tf_i \times idf_i$, we have
$v_1 = [2 \times 0, 1 \times 0.3, 0 \times 0.3] = [0, 0.3, 0]$
$v_2 = [2 \times 0, 0 \times 0.3, 1 \times 0.3] = [0, 0, 0.3]$
So $cos\_sim(D1, D2) = 0$.

(c) $tf_3 = [0, 1, 1]$. Now
$idf_{woof} = log\frac{3}{2} = 0.176$,
$idf_{meow} = log\frac{3}{2} = 0.176$,
$idf_{squeak} = log\frac{3}{2} = 0.176$,
Then
$v_1 = [2 \times 0.176, 1 \times 0.176, 0 \times 0.176] = [0.352, 0.176, 0]$
$v_2 = [2 \times 0, 0 \times 0.3, 1 \times 0.3] = [0.352, 0, 0.176]$
So $cos\_sim(D1, D2) = 0.8$.

(2) [1 point] Naive Bayes and smoothing
Suppose we had 10 restaurant reviews

great food (labeled +)
great food (labeled +)
great food (labeled +)
great food (labeled +)
great food (labeled +)
terrible food (labeled -)
terrible food (labeled -)
terrible food (labeled -)
terrible food (labeled -)
terrible food served (labeled -)

(a) Using [Bernoulli] Naive Bayes without smoothing, compute
P(+ | "great food served") and P(- | "great food served")
(b) Which probability would be larger if you used Laplace (add-one) smoothing?

Solution:

(a) $P(+|\text{"}great\ food\ served\text{"}) = \frac{P(\text{"}great\ food\ served\text{"}|+) \times P(+)}{P(\text{"}great\ food\ served\text{"})}$
$= \frac{P(\text{"}great\text{"}|+) \times P(\text{"}food\text{"}|+) \times P(\text{"}served\text{"}|+) \times P(+)}{P(\text{"}great\text{"}) \times P(\text{"}food\text{"}) \times P(\text{"}served\text{"})}$
$= \frac{\frac{5}{5} \times \frac{5}{5} \times \frac{0}{5} \times \frac{5}{10}}{\frac{5}{10} \times \frac{10}{10} \times \frac{1}{10}}$
$= 0$
Likewise,
$P(-|\text{"}great\ food\ served\text{"}) = \frac{P(\text{"}great\text{"}|-) \times P(\text{"}food\text{"}|-) \times P(\text{"}served\text{"}|-) \times P(-)}{P(\text{"}great\text{"}) \times P(\text{"}food\text{"}) \times P(\text{"}served\text{"})}$
$= \frac{\frac{0}{5} \times \frac{5}{5} \times \frac{1}{5} \times \frac{5}{10}}{\frac{5}{10} \times \frac{10}{10} \times \frac{1}{10}}$
$= 0$

(b) Using Laplace smoothing,
$P(\text{"}great\text{"}|+) = \frac{6}{7}$, $P(\text{"}great\text{"}|-) = \frac{1}{7}$
$P(\text{"}food\text{"}|+) = \frac{6}{7}$, $P(\text{"}food\text{"}|-) = \frac{6}{7}$
$P(\text{"}served\text{"}|+) = \frac{1}{7}$, $P(\text{"}served\text{"}|-) = \frac{2}{7}$
$P(\text{"}terrible\text{"}|+) = \frac{1}{7}$, $P(\text{"}terrible\text{"}|-) = \frac{6}{7}$
Then,
$P(+|\text{"}great\ food\ served\text{"}) = \frac{\frac{6}{7}\times\frac{6}{7}\times\frac{1}{7}\times(1-\frac{1}{7})\times\frac{5}{10}}{\frac{5}{10}\times\frac{10}{10}\times\frac{1}{10}} = 0.9$
$P(-|\text{"}great\ food\ served\text{"}) = \frac{\frac{1}{7}\times\frac{6}{7}\times\frac{2}{7}\times(1-\frac{6}{7})\times\frac{5}{10}}{\frac{5}{10}\times\frac{10}{10}\times\frac{1}{10}} = 0.05$
So the probability of positive review would be larger.