Text classification
Mob. Cloethona.c
Dept of CSE

Doc	text	class
1	I loved the movie	+
2	I hated the movie	-
3	a great movie_ good movie	+
4	poor acting	-
5	great acting a good movie	+

note: Number of unique words, N distinct words
= |vocabulary|

→ convert into feature sets of possible words

→ Number of times a word occurs

10 unique words.

Doc	I	loved	the	movie	hated	a	great	poor	acting	good	class
1	1	1	1	1							+
2	1		1	1	1						-
3				2		1	1			1	+
4								1	1		-
5				1		1	1		1	1	+

Positions ← all word positions in Doc that contain tokens found in vocabulary.

1) probabilities of $+/-$

$$P(+) = \frac{3}{5} = 0.6$$
$$P(-) = \frac{2}{5} = 0.4$$

2) Let n be the no of words in category $(+/-)$

3) Let n_k the no of words occure in the case $(+)$

4) Let $P(w_k|+) = \dfrac{n_{k+1}}{n + |vocabulary|}$ → to make probability less than 1 → this term neg coldp

n_{k+1} → word count

$n + |vocabulary|$ = total no of words in each category

$(+/-)$: count ; $N \rightarrow$ words
$(+/-)$: count ; $N \rightarrow$ no of words
$+$: count ; N no of
coll -; 6

Mrs. chidhane, dept of CSE

1) $P(I|+) = \dfrac{|H|}{14+10} = 0.0833$ 6) $P(loved|+) = \dfrac{|H|}{14+10} = 0.0833$

$P(I|-) = \dfrac{|H|}{6+10} = 0.125$ $P(loved|-) = \dfrac{0+1}{6+10} = 0.0625$

2) $P(The|+) = \dfrac{|H|}{14+10} = 0.0833$ 7) $P(movie|+) = \dfrac{4+1}{14+10} = 0.2083$

$P(The|-) = \dfrac{|H|}{6+10} = 0.125$ $P(movie|-) = \dfrac{|H|}{6+10} = 0.125$

3) $P(a|+) = \dfrac{2+1}{14+10} = 0.125$ 8) $P(great|+) = \dfrac{2+1}{14+10} = 0.125$

$P(a|-) = \dfrac{0+1}{6+10} = 0.0625$ $P(great|-) = \dfrac{0+1}{6+10} = 0.0625$

4) $P(acting|+) = \dfrac{|H|}{14+10} = 0.0833$ 9) $P(good|+) = \dfrac{2+1}{14+10} = 0.125$

$P(acting|-) = \dfrac{|H|}{6+10} = 0.125$ $P(good|-) = \dfrac{0+1}{6+10} = 0.0625$

5) $P(hated|+) = \dfrac{0+1}{14+10} = 0.0417$ 10) $P(poor|+) = \dfrac{0+1}{14+10} = 0.0417$

$P(hated|-) = \dfrac{|H|}{6+10} = 0.125$ $P(poor|-) = \dfrac{|H|}{6+10} = 0.125$

$V_{NB} = \underset{v_j \in V}{argmax}\ P(v_j) \ \pi\ P(a_i|v_j)$

$V_S = +, -$

Instance:
"I hated the poor acting"

$P(y_j|X) = P(+) * P(I|+) * P(The|+) * P(love|+) * P(hated|+) * P(The|+) * P(poor|+) * P(acting|+)$

$= 0.6 * 0.0833 * 0.0833 * 0.0417 * 0.0417 * 0.0833 * 0.0417 * 0.0833$

$= 6.03 * 10^{-7}$

$P(No|X) = P(-) * P(I|-) * P(The|-) * P(loved|-) * P(hated|-) * P(The|-) * P(poor|-) * P(acting|-)$

$= 0.4 * 0.125 * 0.125 * 0.0625 * 0.125 * 0.125 * 0.125 * 0.125$

$= 1.22 * 10^{-5}$

$P(+|X)\ \cancel{X}\ P(-|X)$

$6.03 * 10^{-7} < 1.22 * 10^{-5}$

$\therefore$ document belongs to the class $-$