Machine Learning

Classifiers.

Bag of words.

Docs	f and	f boxing	f film	f great	f plot	f satire	f scenes	f twists	f worst
worst boxing scenes	0	1	0	0	0	0	1	0	1
satire and great plot twists	1	0	0	1	1	1	0	1	0
great scenes great films.	0	0	1	2	0	0	1	0	0

Pros

* Easy
*

Cons

* Creates spare matrix.

  Suppose you have a feature vector of 12000 words. Now on average sentence has 20 words max. so we are creating a matrix of 11980 zeros!

* loss of sequence
* loss of efficiency

Naive Bayes classifier with and ex,

Tr. Data

+ Sorry I'll call later
+ U can call me now
− U have won callnow !!

Te. Data

'Sorry! U cannot unsubscribe

	f1	f'll	f9	f sorry	fU	f call	f can	f have	f later	f me	f now	f won
+	0	1	1	1	0	1	0	0	1	0	0	0
+	0	0	0	0	1	1	1	0	0	1	1	0
−	2	0	0	0	1	1	0	1	0	0	1	1
	1	0	0	1	1	1	1	0	0	0	0	0

smoothing ⇒ $\left(\frac{1+\text{count}}{m+\text{total feature}}\right)$ → This is to remove the occurance of zero

$$P(+|0) = P(+)\,P(\text{sorry}|+)\,P(U|+)\,P(!|+)\,P(\text{can}|+)$$

$$= \frac{2}{3} \times \frac{1+1}{10+12} \times \frac{1+1}{10+12} \times \frac{1}{22} \times \frac{1+1}{22} = 0.000023.$$

$$P(-|0) = P(-)\,P(\text{sorry}|-)\,P(U|-)\,P(!|-)\,P(\text{can}|-)$$

$$= \frac{1}{3} \times \frac{1}{7+12} \times \frac{1+1}{7+12} \times \frac{2+1}{7\times12} \times \frac{1}{7\times12} = 0.000015.$$

So, the test will go to the "positive" class.