NAIVE BAYES CLASSIFIER.

1) Queremos saber si un comentario pertenece a la clase China o no.

	docID	words in document	in class China? (C)
Training set	1	Chinese Beijing Chinese	yes
	2	Chinese Chinese Shanghai	yes
	3	Chinese Macao	yes
	4	Tokyo Japan Chinese	no
Test set	5	Chinese Chinese Chinese Tokyo Japan	?

2) Queremos saber si el caiunto test (5) es + o -.

$$\hat{C} = \max Pr(c) \prod_{i=1}^{n} Pr(f_i, c).$$

3) Calculamos la probabilidad a priori.

$$Pr(c) = \frac{\text{nº documentos de la clase China}}{\text{nº total de documentos in training set}} = \frac{N_c}{N_{docs}}$$

$P(yes) = 3/4$, $P(no) = 1/4$.

4) Calcular las probabilidades condicionales

$$P(t/c) = \frac{T_{ct} + 1}{T'_{CT} + B}$$

T_{CT} = nº ocurrencias del termino t en la clase o grupo C. (+ entrenamiento)

T'_{CT} = nº total de terminos en cada clase (training)

B = nº total de terminos diferentes en todo el grupo entrenamiento
↳ nº palabras diferentes.

$T_{yes\_t} = 8$
$T'_{no\_t} = 3$.
$B = 6$.

$$P(Chinese/yes) = \frac{5+1}{8+6} = (3/7)$$

$$P(Tokyo/yes) = P(Japan/yes) = \frac{0+1}{8+6} = 1/14$$

$$P(Chinese/no) = \frac{1+1}{3+6} = 2/9$$

$$P(Tokyo/no) = P(Japan/no) = \frac{1+1}{3+6} = 2/9$$

5) Multiplicamos la probabilidad a priori de los documentos de cada clase por las probabilidades condicionales de las palabras incluidas en la consulta de test.

$$P(yes/\underline{test}\ doc) = Prob.Priori\ (Chinese^3 \times Japan \times Tokyo)$$
$$= 3/4 \cdot (3/7)^3 \cdot 1/14 \cdot 1/14 = [0,0003]$$

$$P(no/test\ doc) = 1/4 \cdot (2/9)^3 \cdot 2/9 \cdot 2/9 = [0,0006]$$

Sol: Se asigna el documento 5 a: