Exercise _ Text Classification_ Naïve Bayes_ big data

1.- Queremos entender si un comentario pertenece a la clase China o no (Análisis de sentimiento consite en la asiganción de un nuevo comentario a las clases positivas o negativas).

Anteriormente hemos entrenado nuestro sistema, y sabemos qué expresiones se consideran referidas a China o no. Nuestro léxico se muestra en la siguiente tabla

► **Table 13.1** Data for parameter estimation examples.

	docID	words in document	in $c = China$?
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	?

P(yes) = nºyes / nºtotal → 3/4

P(No) = Nº no / nºtotal → 1/4

Queremos saber si el conjunto test es positivo o negativo. Para eso, vamos a utilizar el claificador denominado Naïve Bayes binomial

$$\hat{c} = \max_{c \in C} \Pr(c) \prod_{i=1}^{n} \Pr(f_i \mid c)$$

2.- **Calcularlas probabilidades a priori**

$$\Pr(c_i) = \frac{\text{number of docs of class } c}{\text{total number of docs in training dataset}} = \frac{N_c}{N_{docs}}$$

P(yes)= ¾

P(no)= 1 / 4

3.- Calculamos las probabilidades condicionales

$$\hat{P}(t|c) = \frac{T_{ct}+1}{\sum_{t' \in V}(T_{ct'}+1)} = \frac{T_{ct}+1}{(\sum_{t' \in V} T_{ct'}) + B'}$$

Nº veces término en el grupo de entrenamiento + 1 de una clase

Nº total de términos en cada clase + Nº total de términos en todo el grupo de entrenamiento

- **T_{ct}**= número de ocurrencias del termino t en el grupo de entrenamiento de los documentos clase c c
- **T '$_{ct}$**= número total de términos en cada clase

T '$_{yes-t}$= 8 → mirar tabla.

T '$_{no-t}$= 3 mirar tabla

- **B**= número total de términos diferentes en todo el grupo de entrenamiento (vocabulario)
  - **B=** 6.

Chinese, Beijing, Shangai, Macao, Tokyo, Japan
① ② ③ ④ ⑤ ⑥

¿cuantas veces aparece la palabra chinese en la tabla? → 5

- P(Chinese/yes) = (5+1)/(8+6)= 3/7

↳ T'yes-t

- P(Tokyo/yes)=P(Japan/yes)**= (**0 **+1)/(** 8 **+** 6 **)=**1/14

veces que estas palabras tienen yes

- P(Chinese/no)=(1+1)/(3+6)=2/9

Veces chinese no

- P(Tokyo/no)=P(Japan/no)**=(**1 **+1)/(** 3 **+** 6 **)=**2/9

Veces Tokyo no / Japan no

4.- Multiplicamos la **probabilidad a priori** de los documentos de cada clasepor las **probabilidades condicionales de las palabras incluidas en la consulta** de test

P(yes/ test document)= ¾ * (3/7)^3 *1/14 *1/14= 0.0003

P(no/test document)= $1/4 \cdot (2/9)^3 \cdot 2/9 \cdot 2/9 = 0'0001$

Entonces, ¿el clasificador asigna el documento 5 a ? Chinese yes