Paper Code	Examiner	Department	Office
INT309	Fangyu Wu	Department of Intelligence Science	SD555

Based on the data in Table 1 below:

	docID	contents	in c = China?
Training Set	1	Taipei Taiwan	yes
	2	Macao Taiwan Shanghai	yes
	3	Japan Sapporo	no
	4	Sapporo Osaka Taiwan	no
Test Set	5	Taiwan Taiwan Sapporo	?

Table Q1

1. Estimate a multinomial Naive Bayes classifier, and apply the classifier to the test document.
2. Estimate a Bernoulli Naive Bayes classifier and apply the classifier to the test document.
3. Compute the tf-idf vector (normalized) representations of the documents in the table below. By using Rocchio Classication Algorithm to determine the class of the document.

End of paper

Multinomial:

$$P(c|d) \propto P(c)\Pi_{1\leqslant k\leqslant N_d}P(t_k|c)$$
$$P(t_k|c) = \frac{T_{ct}+1}{\sum_{t'\in V(T_{ct'}+1)}}$$

Bernoulli:

$$P(c|d) \propto P(c)\Pi_{t_k\in Q}P(t_k|c)\Pi_{t_k\notin Q}[1-P(t_k|c)] \quad (5)$$
$$P(t_k|c) = \frac{df_{ct}+1}{N_c+Numeroftclasses} \quad (6)$$

	docID	contents	in c = China?
Training Set	1	Taipei Taiwan	yes
	2	Macao Taiwan Shanghai	yes
	3	Japan Sapporo	no
	4	Sapporo Osaka Taiwan	no
Test Set	5	Taiwan Taiwan Sapporo	?

Table Q1

Taipei, Taiwan, Macao, Shanghai, Japan, Sapporo, Osaka.

1. $\hat{P}(c) = \frac{2}{4} = \frac{1}{2}$

$\hat{P}(\text{Taiwan}|c) = \frac{2+1}{5+7} = \frac{1}{4}$

$\hat{P}(\text{Sapporo}|c) = \frac{0+1}{5+7} = \frac{1}{12}$

$\therefore \hat{P}(c|d_5) = \hat{P}(c)\cdot(\hat{P}(\text{Taiwan}|c))^2\cdot\hat{P}(\text{Sapporo}|c)$

$= \frac{1}{2}\cdot\frac{1}{16}\cdot\frac{1}{12}$

$\cong 0.0026.$

$\hat{P}(\bar{c}) = \frac{2}{4} = \frac{1}{2}$

$\hat{P}(\text{Taiwan}|\bar{c}) = \frac{1+1}{5+7} = \frac{1}{6}$

$\hat{P}(\text{Sapporo}|\bar{c}) = \frac{2+1}{5+7} = \frac{1}{4}$

$\therefore \hat{P}(\bar{c}|d_5) = \hat{P}(\bar{c})(\hat{P}(\text{Taiwan}|c))^2\cdot\hat{P}(\text{Sapporo}|c)$

$= \frac{1}{2}\cdot\frac{1}{36}\cdot\frac{1}{4}$

$\cong 0.0035$

$\because \hat{P}(c|d_5) < \hat{P}(\bar{c}|d_5)$ $\therefore$ Test document doesn't belong to c.

Multinomial:

$$P(c|d) \propto P(c)\Pi_{1\leqslant k\leqslant N_d}P(t_k|c)$$
$$P(t_k|c) = \frac{T_{ct}+1}{\sum_{t'\in V(T_{ct'}+1)}}$$

Bernoulli:

$$P(c|d) \propto P(c)\Pi_{t_k\in Q}P(t_k|c)\Pi_{t_k\notin Q}[1-P(t_k|c)] \quad (5)$$
$$P(t_k|c) = \frac{df_{ct}+1}{N_c+Numerofclasses} \quad (6)$$

	docID	contents	in c = China?
Training Set	1	Taipei Taiwan	yes
	2	Macao Taiwan Shanghai	yes
	3	Japan Sapporo	no
	4	Sapporo Osaka Taiwan	no
Test Set	5	Taiwan Taiwan Sapporo	?

Table Q1

Taipei, Taiwan, Macao, Shanghai, Japan, Sapporo Osaka.

2. $\hat{P}(c) = \frac{2}{4} = \frac{1}{2}$ $\hat{P}(\bar{c}) = \frac{2}{4} = \frac{1}{2}$

$\hat{P}(Taiwan|c) = \frac{2+1}{2+2} = \frac{3}{4}$

$\hat{P}(Sapporo|c) = \frac{0+1}{2+2} = \frac{1}{4}$.

$\hat{P}(Taipei|c) = \hat{P}(Macao|c) = \hat{P}(Shanghai|c) = \frac{1+1}{2+2} = \frac{1}{2}$

$\hat{P}(Japan|c) = \hat{P}(Osaka|c) = \frac{0+1}{2+2} = \frac{1}{4}$.

$\therefore \hat{P}(c|d_5) = \hat{P}(c) \cdot \prod_{t_k\in Q} \hat{P}(t_k|c) \cdot \prod_{t_k\notin Q} (1-\hat{P}(t_k|c))$

$= \frac{1}{2} \times \frac{3}{4} \times \frac{1}{4} \times (1-\frac{1}{2})^3 \times (1-\frac{1}{4})^2$

Taiwan Sapporo Taipei, Macao, SH Japan. Osaka

≈ 0.0066

	docID	contents	in c = China?
Training Set	1	Taipei Taiwan	yes
	2	Macao Taiwan Shanghai	yes
	3	Japan Sapporo	no
	4	Sapporo Osaka Taiwan	no
Test Set	5	Taiwan Taiwan Sapporo	?

Table Q1

$\hat{P}(Taiwan|\bar{c}) = \frac{1+1}{2+2} = \frac{1}{2}$

$= \hat{P}(Japan|\bar{c}) = \hat{P}(Osaka|\bar{c})$

$\hat{P}(Sapporo|\bar{c}) = \frac{2+1}{2+2} = \frac{3}{4}$

$\hat{P}(Taipei|\bar{c}) = \hat{P}(Macao|\bar{c}) = \hat{P}(Shanghai|\bar{c}) = \frac{0+1}{2+2} = \frac{1}{4}$.

$$\therefore \hat{P}(\bar{c}|d_5) = \hat{P}(\bar{c}) \cdot \prod_{t_k \in Q} \hat{P}(t_k|\bar{c}) \cdot \prod_{t_k \notin Q} (1 - \hat{P}(t_k|\bar{c}))$$

$$= \frac{1}{2} \times \frac{1}{2} \times \frac{3}{4} \times \left(1 - \frac{1}{4}\right)^3 \times \left(1 - \frac{1}{2}\right)^2$$

Taiwan Sapporo Tapei, Macao, SH Japan. Osaka.

$$\approx 0.0198$$

$\because \hat{P}(c|d_5) < \hat{P}(\bar{c}|d_5)$

$\therefore$ Test document doesn't belong to c.

3.

	docID	contents	in c = China?
Training Set	1	Taipei Taiwan	yes
	2	Macao Taiwan Shanghai	yes
	3	Japan Sapporo	no
	4	Sapporo Osaka Taiwan	no
Test Set	5	Taiwan Taiwan Sapporo	?

Table Q1

$$idf_t = \log_{10} \frac{N}{df_t}$$

$N = 4$

raw-tf.

term	doc1	doc2	doc3	doc4	doc5	df	idf
Taipei	1	0	0	0	0	1	0.6
Taiwan	1	1	0	1	2	3	0.1
Macao	0	1	0	0	0	1	0.6
Shanghai	0	1	0	0	0	1	0.6
Japan	0	0	1	0	0	1	0.6
Sapporo	0	0	1	1	1	2	0.3
Osaka	0	0	0	1	0	1	0.6

tf-idf (unnormalized)

term	doc1	doc2	doc3	doc4	doc5
Taipei	0.6	0	0	0	0
Taiwan	0.1	0.1	0	0.1	0.2
Macao	0	0.6	0	0	0
Shanghai	0	0.6	0	0	0
Japan	0	0	0.6	0	0
Sapporo	0	0	0.3	0.3	0.3
Osaka	0	0	0	0.6	0

$$C \begin{cases} \overrightarrow{doc1}_N = \dfrac{(0.6,0.1,0,0,0,0,0)}{\sqrt{0.6^2+0.}} = (0.99,0.16,0,0,0,0,0) \\ \overrightarrow{doc2}_N = \dfrac{(0,0.1,0.6,0.6,0,0,0)}{\sqrt{0.1^2+0.6^2+0.6^2}} = (0,0.12,0.7,0.7,0,0,0) \end{cases}$$

$$\bar{C} \begin{cases} \overrightarrow{doc3}_N = \dfrac{(0,0,0,0,0.6,0.3,0)}{\sqrt{0.6^2+0.3^2}} = (0,0,0,0,0.89,0.45,0) \\ \overrightarrow{doc4}_N = \dfrac{(0,0.1,0,0,0,0.3,0.6)}{\sqrt{0.1^2+0.3^2+0.6^2}} = (0,0.15,0,0,0,0.44,0.88) \end{cases}$$

$$\overrightarrow{doc5}_N = \frac{(0,0.2,0,0,0,0.3,0)}{\sqrt{0.2^2+0.3^2}} = (0,0.55,0,0,0,0.83,0).$$

The centroid of the relevant documents is: (class c)

$$c = \frac{\overrightarrow{doc1N} + \overrightarrow{doc2N}}{2} = (0.50, 0.14, 0.35, 0.35, 0, 0, 0)$$

The centroid of the non-relevant documents is: (class $\bar{c}$)

$$\bar{c} = \frac{\overrightarrow{doc3N} + \overrightarrow{doc4N}}{2} = (0, 0.08, 0, 0, 0.45, 0.45, 0.44)$$

The distances from the query to the centroids are:

$$d_{q,c} = \|(0, 0.55, 0, 0, 0, 0.83, 0) - (0.5, 0.14, 0.35, 0.35, 0, 0, 0)\|$$

$$= \|(-0.5, 0.41, -0.35, -0.35, 0, 0.83, 0)\|$$

$$= \sqrt{0.5^2 + 0.41^2 + 0.35^2 \times 2 + 0.83^2} \simeq 1.16$$

$$d_{q,\bar{c}} = \|(0, 0.55, 0, 0, 0, 0.83, 0) - (0, 0.08, 0, 0, 0.45, 0.45, 0.44)\|$$

$$= \|(0, 0.47, 0, 0, -0.45, 0.38, -0.44)\|$$

$$= \sqrt{0.47^2 + 0.45^2 + 0.38^2 + 0.44^2} \simeq 0.87$$

$\because d_{q,c} > d_{q,\bar{c}}$

$\therefore$ test document doesn't belong to c.