Question 1

Based on the data in Table

(i) Classify test document using Multinomial Naive Bayes Classifier

(ii) Classify test document using Bernoulli NB classifier,

(iii) Classify test document using Rocchio classifier

(Use Euclidean distance)

	docID	words in document	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	?

Q 1 Solution

- Multinomial NB
- P (C | d5) = P(C) P(Taiwan | C) P(Taiwan | C) P(Sapporo| C)

  = 0.5 *0.25*0.25*1/12 = 0.0026
- P (Not C | d5) = P(Not C) P(Taiwan | Not C) P(Taiwan | Not C) P(Sapporo| Not C)

  = 0.5 *1/6*1/6*0.25 = 0.00347

d5 is classified in Class Not China

Q 1 Solution

- Bernoulli NB
- P (C | d5) = P(C) P(Taiwan | C)
  P(Sapporo| C) (1 – P(Taipei | C))
  (1 –P(Macao| C)) (1 – P(Taipei | C))
  (1 – P(Shanghai | C)) (1 – P(Japan| C))
  (1 – P(Osaka | C))

= 0.00659

Q 1 Solution

- Bernoulli NB
- P (Not C | d5) = P(Not C) P(Taiwan | Not C)
  P(Sapporo| Not C) (1 – P(Taipei | Not C))
  (1 –P(Macao| Not C)) (1 – P(Taipei | Not C))(1 – P(Shanghai | Not C)) (1 – P(Japan| Not C)) (1 – P(Osaka | Not C))

= 0.0197

Q 1 Solution

- Rocchio

	Taiwan	Sapporo	Japan	Taipei	Osaka	Shanghai	Macao
d_1	0.4	0	0	2	0	0	0
d_2	0.4	0	0	0	0	2	2
d_3	0	1	2	0	0	0	0
d_4	0.4	1	0	0	2	0	0
Centriod of C	0.4	0	0	1	0	1	1
Centriod of Not C	0.2	1	1	0	1	0	0
d_5	0.8	1	0	0	0	0	0

Distance of d5 to centriod of C = 2.36
Distance of d5 to centriod of Not C = 1.61
D5 is classified in class Not C

Question 2

(i) Compute the “export”/POULTRY contingency table for the “Kyoto”/JAPAN in the collection given below.

(ii) Make up a contingency table for which MI is 0 – that is, term and class are independent of each other.

	$e_c = e_{poultry} = 1$	$e_c = e_{poultry} = 0$
$e_t = e_{\text{EXPORT}} = 1$	$N_{11} = 49$	$N_{10} = 27{,}652$
$e_t = e_{\text{EXPORT}} = 0$	$N_{01} = 141$	$N_{00} = 774{,}106$

Collection:

	docID	words in document	in c = Japan?
training set	1	Kyoto Osaka Taiwan	yes
	2	Japan Kyoto	yes
	3	Taipei Taiwan	no
	4	Macao Taiwan Shanghai	no
	5	London	no

Question 3

- Consider the following frequencies for the class *coffee for four terms in the total 100,000 d*ocuments

term	N_{00}	N_{01}	N_{10}	N_{11}
brazil	98,012	102	1835	51
council	96,322	133	3525	20
producers	98,524	119	1118	34
roasted	99,824	143	23	10

- Calculate *mutual information* for the term brazil

Q 2 Solution

Part 1

N11 = 2	N10 = 0
N01 = 0	N00 = 3

Part 2: N11 = N01 and N10 = N00; term is present in half docs of each class

N11 = 3	N10 = 5
N01 = 3	N00 = 5

OR N01 = 0 and N00 = 0 , term is present in all docs in all classes

N11 = 2	N10 = 3
N01 = 0	N00 = 0

Q 3 Solution

N11 = 51	N10 = 1835
N01 = 102	N00 = 98,012

N = 100,000

$$I(U;C) = \frac{N_{11}}{N}\log_2\frac{NN_{11}}{N_{1.}N_{.1}} + \frac{N_{01}}{N}\log_2\frac{NN_{01}}{N_{0.}N_{.1}} + \frac{N_{10}}{N}\log_2\frac{NN_{10}}{N_{1.}N_{.0}} + \frac{N_{00}}{N}\log_2\frac{NN_{00}}{N_{0.}N_{.0}}$$

$$\frac{51}{100000}\log\frac{51 * 100000}{(51 + 1835)(51 + 102)} + \frac{102}{100000}\log\frac{102 * 100000}{(102 + 98012)(51 + 102)} +$$

$$\frac{1835}{100000}\log\frac{1835 * 100000}{(51 + 1835)(1835 + 98012)} + \frac{98012}{100000}\log\frac{98012 * 100000}{(98012 + 1835)(102 + 98012)}$$