	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	?

Table 13.10 Data for parameter estimation exercise.

13.8 Exercises

Exercise 13.1

Which of the documents in Table 13.9 have identical and different bag of words representations for (a) the binomial model (b) the multinomial model?

Exercise 13.2

The rationale for the positional independence assumption is that there is no useful information in the fact that a word occurs in position k of a document. Find exceptions. Consider formulaic documents with a fixed document structure.

Exercise 13.3

The class priors in Figure 13.3 are computed as the fraction of *documents* in the class as opposed to the fraction of *tokens* in the class. Why?

Exercise 13.4

Why is $|C||V| < |D|L_d$ in Table 13.2 expected to hold for most text collections?

Exercise 13.5

Table 13.3 gives binomial and multinomial estimates for the word the. Explain the difference.

Exercise 13.6

Based on the data in Table 13.10, (i) estimate a multinomial Naive Bayes classifier, (ii) apply the classifier to the test document, (iii) estimate a binomial Naive Bayes classifier, (iv) apply the classifier to the test document.

Exercise 13.7

Your task is to classify words as English or not English. Words are generated by a source with the following distribution:

event	word	English?	probability
1	ozb	no	4/9
2	uzu	no	4/9
3	zoo	yes	1/18
4	bun	yes	1/18

(i) Compute the parameters (priors and conditionals) of a multinomial Naive Bayes classifier that uses the letters b, n, o, u, and z as features. Assume a training set that